01010
01010
01010
information

AI AND THE SINGULARITY

A FALLACY OR A GREAT OPPORTUNITY?

Edited by

Robert K. Logan and Adriana Braga

Printed Edition of the Special Issue Published in *Information*

www.mdpi.com/journal/information

MDPI

AI AND THE SINGULARITY

AI AND THE SINGULARITY

A FALLACY OR A GREAT OPPORTUNITY?

Special Issue Editors

Robert K. Logan
Adriana Braga

MDPI • Basel • Beijing • Wuhan • Barcelona • Belgrade • Manchester • Tokyo • Cluj • Tianjin

Special Issue Editors

Robert K. Logan
University of Toronto
Canada

Adriana Braga
Pontifícia Universidade Católica
do Rio de Janeiro (PUC-RJ)
Brazil

Editorial Office
MDPI
St. Alban-Anlage 66
4052 Basel, Switzerland

This is a reprint of articles from the Special Issue published online in the open access journal *Information* (ISSN 2078-2489) (available at: https://www.mdpi.com/journal/information/special_issues/AI%26Singularity).

For citation purposes, cite each article independently as indicated on the article page online and as indicated below:

LastName, A.A.; LastName, B.B.; LastName, C.C. Article Title. *Journal Name* **Year**, *Article Number*, Page Range.

ISBN 978-3-03936-483-1 (Hbk)
ISBN 978-3-03936-484-8 (PDF)

Contents

About the Special Issue Editors

Robert K. Logan, Prof., (Ph.D. MIT 1965) is an Emeritus Professor of physics, a fellow of St. Michael's College, and a member of the School of the Environment at the University of Toronto. He is also the Chief Scientist at OCAD (Ontario College of Art and Design) University in the Strategic Innovation Lab. He has a variety of experiences as an academic and has engaged in research in a variety of fields, including physics, media ecology, complexity theory, information theory, systems biology, environmental science, linguistics, and industrial design. He has published with and collaborated with Marshall McLuhan and continues his McLuhan studies research. He is also an author or editor of 19 books, 23 book chapters, and over 150 articles in refereed journals. He once taught The Poetry of Physics and the Physics of Poetry course at the U. of Toronto and has published a book with the same title as that of the course. He was the recipient of the Walter J. Ong Award for Career Achievement in Scholarship by the Media Ecology Association.

Adriana Braga is an Associate Professor of the Graduate Program of Social Communication at the Pontifical Catholic University of Rio de Janeiro (PUC-Rio), researcher of the National Council for Scientific and Technological Development (CNPq/Brazil), coordinator of the Research Group on Digital Interactions (GRID) and is a Visiting Professor at the University of Macau, China. With a Ph.D. in Communication Sciences and BA in Psychology, she is also the author of the books *Introdução à Ecologia das Mídias* (*Introduction to Media Ecology* with Lance Strate and Paul Levinson, 2019); *Personas Materno-Eletrônicas* (*Maternal-Electronic Personae*, 2008); *CMC, Identidades e Género* (*CMC, Identities and Gender*, in Portugal, 2005); and *Corpo-Verão* (*Summer-Body*, 2016). Her Ph.D. dissertation won the CAPES Thesis Award (MEC/Brazil) and Harold Innis Award (Media Ecology Association/USA). She serves as Vice President of the Media Ecology Association. She currently coordinates the Digital Media Lab (LabMiD), which develops projects in digital communication, literature, Brazilian culture, mobile telephony, digital literacy, gender, sociolinguistics, and technology.

Preface to ”AI AND THE SINGULARITY”

Dear Readers,

The essays that we have collected address the question of whether the technological singularity—the notion that AI-based computers can exceed human intelligence—is a fallacy or a great opportunity. To address this question, we have invited a group of scholars whose positions on the singularity range from advocates to skeptics. No conclusion can be reached, as the development of artificial intelligence is still in its infancy, and there is much wishful thinking and imagination in this issue rather than trustworthy data. The reader will find a cogent summary of the issues faced by researchers who are working to develop the field of artificial intelligence and, in particular, artificial general intelligence. The only conclusion that can be reached is that there exists a variety of well-argued positions as to where AI research is headed.

We would be happy to hear from you, and we welcome your opinions.

Robert K. Logan, Adriana Braga

Special Issue Editors

information

Editorial

AI and the Singularity: A Fallacy or a Great Opportunity?

Adriana Braga [1] **and Robert K. Logan** [2,*]

[1] Department of Social Communication, Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio), Rio de Janeiro 22451-900, Brazil; adrianabraga@puc-rio.br

[2] Department of Physics, University of Toronto, Toronto, ON M5S 1A7, Canada

* Correspondence: logan@physics.utoronto.ca; Tel.: +1-416-361-5928

Received: 18 February 2019; Accepted: 18 February 2019; Published: 21 February 2019

Abstract: We address the question of whether AI, and in particular the Singularity—the notion that AI-based computers can exceed human intelligence—is a fallacy or a great opportunity. We have invited a group of scholars to address this question, whose positions on the Singularity range from advocates to skeptics. No conclusion can be reached as the development of artificial intelligence is still in its infancy, and there is much wishful thinking and imagination in this issue rather than trustworthy data. The reader will find a cogent summary of the issues faced by researchers who are working to develop the field of artificial intelligence and in particular artificial general intelligence. The only conclusion that can be reached is that there exists a variety of well-argued positions as to where AI research is headed.

Keywords: artificial intelligence; artificial general intelligence; Singularity; cognition; emotion; computers; information; meaning; intuition; wisdom

1. Introduction

We made a call for papers that either support or criticize the lead paper for this Special Issue entitled *The Emperor of Strong AI Has no Clothes: Limits to Artificial Intelligence*. In this lead paper, we argued that the premise of the technological Singularity, based on the notion that computers will one day be smarter than their human creators, is false, and made use of the techniques of media ecology. We also analyzed the comments of other critics of the Singularity, as well as of those supportive of this notion. The notion of intelligence that advocates for the technological Singularity does not take into account the full dimension of human intelligence. They treat artificial intelligence as a figure without a ground. Human intelligence, as we will show, is not based solely on logical operations and computation, but also includes a long list of other characteristics, unique to humans, which is the ground that supporters of the Singularity ignore. The list includes curiosity, imagination, intuition, emotions, passion, desires, pleasure, aesthetics, joy, purpose, objectives, goals, telos, values, morality, experience, wisdom, judgment, and even humor. We asked that the contributors to this special issue either support or critique the thesis we developed in the lead paper. We received contributions concerning both sides of the argument, and have therefore put together an interesting collection of viewpoints that explores the pros and cons of the notion of the Singularity. The importance of this collection of essays is that it explores both the challenges and the opportunities of artificial general intelligence in order to clarify an important matter that has been shadowed by ideological wishful thinking, biased by marketing issues, and influenced by creative imagination about how the future would be, rather than supported by trustworthy data and grounded research.

In this sense, we want to share an article by Daniel Tukelang entitled *10 Things Everyone Should Know About Machine Learning* [1] that he has kindly given us permission to reproduce, in order to give the reader some background on the issues surrounding the notion of the Singularity.

Daniel Tukelang wrote, “As someone who often finds himself explaining machine learning to nonexperts, I offer the following list as a public service announcement”:

1. Machine learning means learning from data; AI is a buzzword. Machine learning lives up to the hype, and there is an incredible number of problems that you can solve by providing the right training data to the right learning algorithms. Call it AI if that helps you sell it, but know that AI, at least as it is used outside of academia, is often a buzzword that can mean whatever people want it to mean.
2. Machine learning is about data and algorithms, but mostly data. There is a lot of excitement about advances in machine learning algorithms, and particularly about deep learning. However, data is the key ingredient that makes machine learning possible. You can have machine learning without sophisticated algorithms, but not without good data.
3. Unless you have a lot of data, you should stick to simple models. Machine learning trains a model from patterns in your data, exploring a space of possible models defined by parameters. If your parameter space is too big, you will overfit to your training data and train a model that does not generalize beyond it. A detailed explanation requires more math, but as a rule, you should keep your models as simple as possible.
4. Machine learning can only be as good as the data you use to train it. The phrase “garbage in, garbage out” predates machine learning, but it aptly characterizes a key limitation of machine learning. Machine learning can only discover patterns that are present in your training data. For supervised machine learning tasks like classification, you will need a robust collection of correctly labeled, richly featured training data.
5. Machine learning only works if your training data is representative. Just as a fund prospectus warns that “past performance is no guarantee of future results”, machine learning should warn that it is only guaranteed to work for data generated by the same distribution that generated its training data. Be vigilant of skews between training data and production data, and retrain your models frequently, so they do not become stale.
6. Most of the hard work for machine learning is data transformation. From reading the hype about new machine learning techniques, you might think that machine learning is mostly about selecting and tuning algorithms. The reality is more prosaic: most of your time and effort goes into data cleansing and feature engineering—that is, transforming raw features into features that better represent the signal in your data.
7. Deep learning is a revolutionary advance, but it is not a magic bullet. Deep learning has earned its hype by delivering advances across a broad range of machine learning application areas. Moreover, deep learning automates some of the work traditionally performed through feature engineering, especially for image and video data. But deep learning is not a silver bullet. You cannot just use it out of the box, and you will still need to invest significant effort in data cleansing and transformation.
8. Machine learning systems are highly vulnerable to operator error. With apologies to the NRA, “Machine learning algorithms don’t kill people; people kill people.” When machine learning systems fail, it is rarely because of problems with the machine learning algorithm. More likely, you have introduced human error into the training data, creating bias or some other systematic error. Always be skeptical, and approach machine learning with the discipline you apply to software engineering.
9. Machine learning can inadvertently create a self-fulfilling prophecy. In many applications of machine learning, the decisions you make today affect the training data you collect tomorrow. Once your machine learning system embeds biases into its model, it can continue generating new training data that reinforce those biases. And some biases can ruin people’s lives. Be responsible: do not create self-fulfilling prophecies.

10. AI is not going to become self-aware, rise up, and destroy humanity. A surprising number of people seem to be getting their ideas about artificial intelligence from science fiction movies. We should be inspired by science fiction, but not so credulous that we mistake it for reality. There are enough real and present dangers to worry about, from consciously evil human beings to unconsciously biased machine learning models. So you can stop worrying about SkyNet and "superintelligence".

There is far more to machine learning than I can explain in a top 10 list. But hopefully, this serves as a useful introduction for nonexperts.

2. Materials and Methods

As the question of whether or not the Singularity can be achieved can only be determined after many years of research, the essays in this collection are based largely on each author's opinions of human cognition and the progress made in the field of artificial intelligence.

3. Results

The basic result of this collection of essays and opinions is an extensive list of the challenges in the development of artificial general intelligence and the possibility of the Singularity—the notion that an artificial general intelligence can be achieved that exceeds human intelligence.

4. Discussion

One of the things that readers should keep in mind when reading the articles that we have collected for this Special Issue is that AI research is still in the early stages, and AI as a data processing device is not foolproof.

The recent detection of the gravitational wave from the merger of two neutron stars as a result of scientists not trusting their AI-configured automated data-processing programs underscores the thesis that one cannot rely on artificial intelligence alone. The case study that is reviewed illustrates the point that AI, combined with human intervention, produces the most desirable results, and that AI by itself will never entirely replace human intelligence.

Ethan Siegel, an astrophysicist, science communicator, and NASA columnist, in a recent article entitled *LIGO's Greatest Discovery Almost Didn't Happen* [2], demonstrated that AI-configured computers will never replace human intelligence, but that they are nevertheless important tools that enhance human intelligence. If the scientists had relied solely on the results of their AI-configured automated data-processing programs, they would have missed a critical observation of the production of gravity waves from the merger of two neutron stars—an extremely rare event never before observed.

There are altogether three observatories for detecting gravity waves, with two LIGO (The Laser Interferometer Gravitational-Wave Observatory) detectors located at Hanford Washington and Livingston Louisiana, and the EGO (European Gravitational Observatory) detector located near Pisa, Italy. The three detectors were in agreement when they observed the first detected gravitational wave that emanated from the merger of two black holes.

A short time after the detection of the first gravitational wave at all three detectors, a signal was received at the Hanford detector consistent with the merger of two neutron stars. The problem was that no signal was registered at the other two detectors, as should have been the case if a gravitational wave had arrived at our planet according to the automated data-processing program in use at the three detectors. Without the corroborating evidence from the two other detectors, the team at Hanford would have been forced to conclude that the signal was not the detection of a gravitational wave but rather a glitch in the system. However, one of the scientists, Reed Essick, decided that it was worth examining the data from the other detectors to determine if there was a signal that had been missed as a result of a glitch at these detectors. He went through the painstaking task of examining every signal that might have been received by the Livingston detector around the time of the event detected at

Hanford. To his delight, he found that a signal had been registered at the Livingston detector but had been overlooked by the automated computer program because of a glitch at that detector. An analysis of the detector at Pisa revealed that, at the time of the event at Hanford, the Pisa detector was in a blind spot for observing the event of the two neutron stars merging. Corroborating data came from NASA's Fermis satellite, which had detected a "short period gamma-ray burst" that arrived two seconds after the arrival of the gravity wave and was consistent with the merger of two neutron stars. As a result of the due diligence and the perseverance of the LIGO team led by Essick, an astronomically important observation was rescued. In his article, Siegel drew the following conclusion from this episode:

> Here's how scientists didn't let it slip away . . . If all we had done was look at the automated signals, we would have gotten just one "single-detector alert," in the Hanford detector, while the other two detectors would have registered no event. We would have thrown it away, all because the orientation was such that there was no significant signal in Virgo, and a glitch caused the Livingston signal to be vetoed. If we left the signal-finding solely to algorithms and theoretical decisions, a 1-in-10,000 coincidence would have stopped us from finding this first-of-its-kind event. But we had scientists on the job: real, live, human scientists, and now we've confidently seen a multimessenger signal, in gravitational waves and electromagnetic light, for the very first time.

Consequently, the conclusion that we reached as a result of Siegel's story and his conclusion is as follows:

1. AI combined with human intervention produces the most desirable results.
2. AI by itself will never entirely replace human intelligence.
3. One cannot rely on AI, no matter how sophisticated, to always get the right answer or reach the correct conclusion.

The variety of opinions presented in this Special Issue will give the readers much to think about vis-à-vis AI, artificial general intelligence, and the Singularity. Readers are invited to correspond to the coeditors with their reactions to, and opinions of, these essays. Thank you for your attention.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Tukelang, D. 10 Things Every Small Business Should Know About Machine Learning. Available online: https://www.inc.com/quora/10-things-every-small-business-should-know-about-machine-learning.html (accessed on 29 November 2017).
2. Siegel, E. LIGO's Greatest Discovery Almost Didn't Happen. Available online: https://medium.com/starts-with-a-bang/ligos-greatest-discovery-almost-didn-t-happen-a315e328ca8 (accessed on 24 April 2018).

information

Article

The Emperor of Strong AI Has No Clothes: Limits to Artificial Intelligence

Adriana Braga [1] and Robert K. Logan [2,*]

[1] Department of Social Communication, Pontifícia Universidade Católica do Rio de Janeiro (PUC-RJ), R. Marquês de São Vicente, 225—Gávea, Rio de Janeiro 22451-900, RJ, Brazil; adrianabraga1@yahoo.com.br

[2] Department of Physics, University of Toronto, 60 St. George, Toronto, ON M5S 1A7, Canada

* Correspondence: logan@physics.utoronto.ca; Tel.: +1-(416)-978-8632

Received: 31 October 2017; Accepted: 22 November 2017; Published: 27 November 2017

Abstract: Making use of the techniques of media ecology we argue that the premise of the technological Singularity based on the notion computers will one day be smarter that their human creators is false. We also analyze the comments of other critics of the Singularity, as well supporters of this notion. The notion of intelligence that advocates of the technological singularity promote does not take into account the full dimension of human intelligence. They treat artificial intelligence as a figure without a ground. Human intelligence as we will show is not based solely on logical operations and computation, but also includes a long list of other characteristics that are unique to humans, which is the ground that supporters of the Singularity ignore. The list includes curiosity, imagination, intuition, emotions, passion, desires, pleasure, aesthetics, joy, purpose, objectives, goals, telos, values, morality, experience, wisdom, judgment, and even humor.

Keywords: technological Singularity; intelligence; emotion; artificial general intelligence; artificial intelligence; computer; logic; figure/ground

---

> *The true sign of intelligence is not knowledge but imagination.* —Albert Einstein

> *Levels of consciousness. Knowing what one knows. Language is necessary for knowing what one knows as one talks to oneself. But computer have no need or desire to communicate with others and hence never created language and without language one cannot talk to oneself and hence computers will never be conscious* [1].

1. Introduction

The notion of the technological singularity or the idea that computers will one day be more intelligent than their human creators has received a lot of attention in recent years. A number of scholars have argued both for and against the idea of a technological singularity using a variety of different arguments. A sample of these opinions for and against the idea of the technological singularity can be found in two collections of short essays, entitled Special Report: The Singularity [2] and *What to Think About Machines that Think* [3] as well as the critical writings of Herbert Dreyfus [4–7]. We will analyze their different positions and make use of their arguments, which we will integrate into our own critiques of both the idea that computers can think, and the idea of the Singularity, or the idea that machines through the use of Artificial General Intelligence (AGI, sometimes referred to simply as AI) can become more intelligent than their human creators. We intend to show that despite the usefulness of artificial intelligence, that the Singularity is an over extension of AI and that no computer can ever duplicate the intelligence of a human being.

We will argue that the emperor of AI is quite naked by exploring the many dimensions of human intelligence that involve characteristics that we believe cannot be duplicated by silicon based forms of intelligence because machines lack a number of essential properties that only a flesh and blood living organism, especially a human, can possess. In short, we believe that artificial intelligence (AI) or its stronger version artificial general intelligence (AGI) can never rise to the level of human intelligence because computers are not capable of many of the essential characteristics of human intelligence, despite their ability to out-perform us as far as logic and computation are concerned. As Einstein once remarked "Logic will get you from A to B. Imagination will take you everywhere".

What motivated us to write this essay is our fear that some who argue for the technological singularity might in fact convince many others to lower the threshold as to what constitutes human intelligence so that it meets the level of machine intelligence, and thus devalue those aspects of human intelligence that we (the authors) hold dear such as imagination, aesthetics, altruism, creativity, and wisdom.

To be a fully realized human intelligent being it is necessary, in our opinion, to have these characteristics. We will suggest that these many aspects of the human experience that are associated uniquely with our species Homo sapiens (wise humans) do not have analogues in the world of machine intelligence, and that as a result the notion that an artificial intelligent machine-based system that is more intelligent than a human is not possible and that the notion of the technological singularity is basically science fiction. We recognize that the attributes that we listed above that constitute what we consider to be intelligence are arrived at subjectively. Perhaps we are defining what we believe is a humane form of intelligence as has been suggested kindly by one of the reviewers of an earlier version of this essay. But, that is one of the objectives of this essay, a desire to make sure that in the desire to gain the benefits of AI, we as a society do not degrade the humaneness of what is considered intelligence. Human intelligence and machine intelligence are of a completely different nature so to claim that one is greater than the other is like comparing the proverbial apples and oranges. They are different and they are both valuable and one should not be mistaken for the other.

There is a subjective, non-rational (or perhaps extra-rational) aspect of human intelligence, which a computer can never duplicate. We do not want to have intelligence as defined by Singularitarians, who are primarily AI specialists and as a result are motivated to exaggerate their field of research and their accomplishments as is the case with all specialists. Engineers should not be defining intelligence. Consider the confusion engineers created by defining Shannon's measure of signal transmission as information (see Braga and Logan [8]).

To critique the idea of the Singularity we will make use of the ideas of Terrence Deacon [9], as developed in his study *Incomplete Nature: How Mind Emerged from Matter*. Deacon's basic idea is that for an entity to have sentience or intelligence it must also have a sense of self [9] pp. 463–484. In his study, Deacon [9] p. 524 defines information "as about something for something toward some end". As a computer or an AI device has no sense of self (i.e., no one is home), it has no information as defined by Deacon. The AI device only has Shannon information, which has no meaning for itself, i.e., the computer is not aware of what it knows as it deals with one bit of data at a time. We will discover that many of the other critiques of the singularity that we will reference parallel our notion that a machine has no sense of self, no objectives or ends for which it strives, and no values.

We will also make use of media ecology and the insights of Marshall McLuhan [10] including:

- the notion that a figure must be understood in the context of the ground or environment in which it operates,
- the notion that every technology brings both service and disservice, and,
- the idea that any technology pushed far enough flips into opposite or complementary form.

Given that the medium is the message, as McLuhan [10] proclaimed, we will examine the medium of the computer and its special use as an artificial intelligence (AI) device with particular attention to strong AI or AGI. Our basic thesis is that computers, together with AI, are a form of technology and a medium that extends human intelligence not a form of intelligence itself.

Our critique of AGI will make use of McLuhan's [10] technique of figure/ground analysis, which is at the heart of his iconic one-liner the "medium is the message" that first appeared in his book *Understanding Media*. The medium independent of its content has its own message. The meaning of the content of a medium, the figure, is affected by the ground in which it operates, the medium itself. The problem that the advocates of AGI and the Singularity make is they regard the computer as a figure without a ground. As McLuhan once pointed out "logic is figure without ground" [11]. A computer is nothing more than a logic device and hence it is a figure without a ground. A human and the human's intelligence are each a figure with a ground, the ground of experience, emotions, imagination, purpose, and all of the other human characteristics that computers cannot possibly duplicate because they have no sense of self.

While we are critical of the notion of the idea of the Singularity, we are quite positive re the value of AI. We also believe, like Rushkoff [12] pp. 354–355, that networked computers will increase human intelligence by allowing humans to network and share their insights. We also concur with Benjamin Bratton [13]:

> "AI can have an enormously positive role to play at an infrastructural level, not just the augmentation of an individual's intelligence, but the augmentation of systemic intelligence and the ability of infrastructural systems to automate what we call political decision or economic decision."

The pattern recognition capabilities of big data will assist humans to make new discoveries, but it will require human intelligence to guide the AI devices as to what patterns to look for. In short, AI guided by human intelligence will always be more productive than AGI working on its own. As pointed out by one of the reviewers of our essay, other forms of a technological singularity that do not try to duplicate human intelligence are altogether possible but they are not the subject of this essay.

2. Origin of the Singularity Idea

The following excerpt from the article, *Technological Singularity* in Wikipedia, accessed 15 September 2017, summarizes the rise of the concept in the early days of the computer age beginning with a conversation between John Von Neumann and Stan Ulam.

> The first use of the term "singularity" in this context was made by Stanislaw Ulam in his 1958 obituary for John Von Neumann, in which he mentioned a conversation with von Neumann about the "ever accelerating progress of technology and changes in the mode of human life, which gives the appearance of approaching some essential singularity in the history of the race beyond which human affairs, as we know them, could not continue". The term was popularized by mathematician, computer scientist and science fiction author Vernor Vinge, who argues that artificial intelligence, human biological enhancement, or brain-computer interfaces could be possible causes of the singularity. Futurist Ray Kurzweil cited von Neumann's use of the term in a foreword to von Neumann's classic *The Computer and the Brain* (https://en.wikipedia.org/wiki/Technological_singularity).

2.1. The Belief in the Singularity

Let us examine the roots of the fantasy that humans will one day be obsolesced by computers. A number of advocates of strong AI or AGI have suggested that in the not too distant futures (2045 according to Ray Kurzweil [14]) programmers will design a computer with an AI or AGI capability that will allow it to design a computer even more intelligent than itself and that computer will be able to do the same and by a process of iteration a technological singularity point will be arrived at where post-Singularity computers will be far more intelligent than us poor humans who only have an intelligence that is designed by nature through natural selection and evolution. At this point, according to those that embrace this idea the super-intelligent computers will take over and we human will become their docile servants.

This scenario is not the product of science fiction writers. Rather, they are the scenarios being that are painted by information and computer scientists who are advocates of strong AI or AGI and the idea of the Technological Singularity.

The idea of the perfectibility of human intelligence can be traced back to the Enlightenment and the encyclopaedist Condorcet who wrote,

> Nature has set no term to the perfection of human faculties; that the perfectibility of man is truly indefinite; and that the progress of this perfectibility, from now onwards independent of any power that might wish to halt it, has no other limit than the duration of the globe upon which nature has cast us (www.historyguide.org/intellect/sketch.html).

It is worth noting that the idea of the Singularity is the product of techno-optimists who have made other predictions in the past that did not pan out. For example, the techno-optimists once predicted that with the efficiency of computers, that there would be a dramatic decrease in the number of hours we would need to work, which we have yet to seen (http://paleofuture.gizmodo.com/the-late-great-american-promise-of-less-work-1561753129). "For instance, computer workstations have revolutionized office and retail ... Yet the dramatic increase in productivity has not led to a shorter work week or to a more relaxed work environment" [15] p. 74. Instead, it has led to companies using the efficiency of computers to reduce the number of workers needed to perform certain tasks.

Techno-optimists also predicted that offices would be practically paper-free, when in fact the amount of paper in offices has actually increased.

> Over the past thirty years, many people have proclaimed the imminent arrival of the paperless office. Yet even the World Wide Web, which allows almost any computer to read and display another computer's documents, has increased the amount of printing done. The use of e-mail in an organization causes an average 40% increase in paper consumption [16] p. 1.

2.2. Singularity Advocates with a Spiritual Dimension to Their Belief

When we first began to think of this project and work on the research that led to this essay the biggest mystery for us was what motivates a tough minded scientific thinker to believe that human intelligence can be programmed into or instantiated on a computer, a mechanical machine. Then, we read Steven Pinker's [17] pp. 5–8 short little essay "Thinking Does Not Imply Subjugating", and in the following few sentences he succinctly described his romance in the belief that human intelligences could be captured in a machine:

> The cognitive feats of the brain can be explained in physical terms ... This is a great idea for two reasons. First, it completes a naturalistic understanding of the universe, exorcising occult souls, spirits and ghosts in the machine. Just as Darwin made it possible for a thoughtful observer of the natural world to do without creationism, Turing and others made it possible for a thoughtful observer of the cognitive world to do so without spiritualism.

However, some proponents of the Singularity have a religious zeal to them not in the theist sense but somewhat similar to the beliefs of the deists. Here is a collection of positions by Singularity zealots that have in our opinion to varying degrees a religious tone to them.

Frank Tipler [18] has an amusing solution for the inevitable fact that once our sun runs out of nuclear fuel and can no longer provide the conditions that make life on Earth sustainable that our only hope for the survival of human culture will be AI computers (AIs) that do not require the conditions that make carbon-based life sustainable. He suggests that the AIs will take to outer space. And

> any human who wants to join the AIs in their expansion can become a human upload, a technology that should be developed about the same time as AI technology . . . If you can't beat 'em, join 'em . . . When this doom is at hand, any human who remains alive and doesn't want to die will have no choice but to become a human upload. The AIs will save us all.

The parallels of Tipler's proposal with Christianity are striking. God Is dead but AI has been born and it is our Savior and like Jesus's self-described appellation, it is "the son of man". AI, not Jesus, "will save us all" and eternal life can be found in an AI computer somewhere in space like the "kingdom of heaven" (Matthew 3.2) and not here on Earth.

Anthony Garrett Lisi [19], in an article entitled "I, for One, Welcome our Machine Overlords", claimed: "Computers share knowledge much more easily than humans do, and they keep that knowledge longer, becoming wiser than humans". Lisi in his attempt to find a higher power makes the mistake that wisdom comes from knowledge. Knowledge is about using information to achieve one's objectives and wisdom is the ability to choose objective consistent with one's values. How can a computer have values? The values of a computer are those of its programmers.

Pamela McCorduck [20] p. 53 in an article entitled "An Epochal Human Event" opined, "We long to preserve ourselves as a species. For all of the imaginary deities that we have petitioned throughout history who have failed to protect us from nature, from one another, from ourselves—we're finally ready to call on our own enhanced, augmented minds instead". Her god is "our own enhanced, augmented minds".

Sam Harris [21] suggests that a super intelligent AGI could achieve 20,000 years of intellectual work in a week. Scientific work requires making observation, designing, and building observational tools. His closing comments reveal what we have identified as the quasi-religious fervor of the AGI advocates: "We seem to be in the process of building a god. Now would be a good time to wonder whether it will (or can be) a good one".

Gregory Paul [22] writes, "The way for human minds to avoid becoming obsolete is to join in the cyber civilization; out of growth-limited bio brains into rapidly improving cyber brains". He then suggests that we can then give up our physical bodies which would then benefit the Earth's biosystem. This is a variation on the Christian idea that we can have everlasting life as pure spirits. For Gregory Paul heaven will be in the clouds, computer clouds.

James Croak [23] suggests that, "Fear of AI is the latest incarnation of our primal unconscious fear of an all-knowing, all-powerful angry God dominating us–but in a new ethereal form".

Douglas Hofstadter [24] provides us with an apocalyptic scenario of the impact of the Singularity, which he believes is a "couple of centuries" away. He suggests that the ramifications "will be enormous, since the highest form of sentient beings on the planet will no longer be human. Perhaps these machines—our 'children'—will be vaguely like us and will have culture similar to ours, but most likely not. In that case, we humans may well go the way of the dinosaurs".

Perhaps the most explicit example of the religious devotion to the idea of the Singularity comes from AI programmer Anthony AI programmer Anthony Levandowski, who is famous for his work on self-driving cars, first for Waymo (a subsidiary of Alphabet, Google's holding company) and later for Uber. Levandowski has founded a non-profit religious organization, the Way of the Future that plans to "develop and promote the realization of a Godhead based on artificial intelligence *and through understanding and worship of the Godhead contribute to the betterment of society* (https://www.wired.com/story/god-is-a-bot-and-anthony-levandowski-is-his-messenger)".

We end this section with John Horgan's [25] humorous and skeptical take on the eternal life belief by Singularity advocates, who believe that humans can be uploaded onto a computer:

> I would love to believe that we are rapidly approaching "the singularity". Like paradise, technological singularity comes in many versions, but most involve bionic brain boosting. At first, we'll become cyborgs, as stupendously powerful brain chips soup up our

> perception, memory, and intelligence and maybe even eliminate the need for annoying TV remotes. Eventually, we will abandon our flesh-and-blood selves entirely and upload our digitized psyches into computers. We will then dwell happily forever in cyberspace . . . Kurzweil says he has adopted an antiaging regimen so that he'll live long enough to live forever.

Horgan remains skeptical of the uploading of human brains on to computers, mainly because neuroscientists have such a sketchy understanding of how the brain operates or how it stores memories, or what are the roles of various chemicals found in the brain.

> Neurotransmitters, which carry signals across the synapse between two neurons, also come in many different varieties. In addition to neurotransmitters, neural-growth factors, hormones, and other chemicals ebb and flow through the brain, modulating cognition in ways that are both profound and subtle [25].

2.3. Have We Become the Servomechanisms of Our Computers?

Although we firmly believe that machine intelligence can never exceed human intelligence, there is still a very real danger, however, that we can lose some of our autonomy to AI or AGI through a decline in what we regard as human intelligence and hence how we view the nature of the human spirit. There are other instances where we have partially lost our autonomy to other technologies. One example is our total dependence on the automobile and the burning of fossil fuels that now threatens our very existence due to global warming and climate change.

In Chapter 4 of *Understanding Media: The Gadget Lover Narcissus as Narcosis,* Marshall McLuhan [10] describes how we allow our technologies to take over and control us so that we become their servo mechanisms. We believe this is an apt description of the AGI computer 'gadget lovers' who support the notion of the inevitability of the technological singularity. McLuhan [10] p. 51 wrote:

> The Greek myth of Narcissus is directly concerned with a fact of human experience, as the word Narcissus indicates. It is from the Greek word narcosis, or numbness. The youth Narcissus mistook his own reflection in the water for another person. This extension (or amplification) of himself by mirror numbed his perceptions until he became the servomechanism of his own extended or repeated image . . . Now the point of this myth is the fact that men at once become fascinated by any extension of themselves (i.e., their technological extensions) . . . Such amplification is bearable by the nervous system only through numbness or blocking of perception . . . To behold, use or perceive any extension of ourselves in form is necessarily to embrace it . . . By continuously embracing technologies, we relate ourselves to them as servo-mechanisms. That is why we must, to use them at all, serve these objects, these extensions of ourselves, as gods or minor religions . . . Physiologically, man in the normal use of technology (or his variously extended body) is perpetually modified by it and in turn finds ever new ways of modifying his technology. Man becomes, as it were, the sex organs of the machine world, as the bee of the plant world, enabling it to fecundate and to evolve ever new forms.

We have quoted this rather long excerpt from McLuhan because it seems to describe some advocates of AGI, who, in our opinion, have become the servomechanisms of their computer technology and even suggest that we will become even more than their servomechanisms, but we will become their slaves. Like Narcissus who fell in love with his own image reflected in the water, the advocates of the technological singularity see a reflection of themselves in the computers that they program. Riffing on the one-liner "garbage in, garbage out" we would suggest "computer worship in, narcissism out".

Quentin Hardy [26], a strong advocate of strong AI, wrote, "we have met the AI and it is us". AI is like the pool that Narcissus looked into and fell in love with his own image. AI is just a reflection of one aspect of our own intelligence (the logical and rational aspect), and we, or some of us, have fallen in love with it. Just as Narcissus was suffering from narcosis because he was so mesmerized by his own reflection he could not get beyond himself. Many of the advocates of AGI or strong AI are mesmerized by the beauty of logic and rationality to such an extent that they dismiss the emotional, the non-rational, the poetry of poiesis, and the emotional side of intelligence. Human intelligence is bicameral. Not a neat division of the analytic/rational left brain and the artistic/intuitive right brain, but a synthesis of these two aspects of the human mind. The AI computer brain is unicameral with a left brain bias. It lacks the neurochemistry, such as dopamine, serotonin, and other agents that are triggered by or are part of human emotional life.

Timothy Taylor [27] introduces the idea of *denkraumverlust,* whereby one confuses the representation of something for the thing itself. He cites the example of the Pygmalion myth where a sculptor falls in love with the figure of a woman that he sculpted. In a similar way, creators of AGI have fallen in love with their creations attributing properties to them that they do not possess like the creator of Pygmalion or the male lead in the movie *Her*.

3. The Ground of Intelligence—What Is Missing in Computers

At the core of our critique of the technological singularity is our belief that human intelligence cannot be exceeded by machine intelligence because the following set of human attributes are essential ingredients of human intelligence, and they cannot, in our opinion, be duplicated by a machine. The most important of these is that humans have a sense of self and hence have purpose, objectives, goals, and telos, as has been described by Terrence Deacon [9] pp. 463–484 in his book *Incomplete Nature*. As a result of this sense of self, humans also have curiosity, imagination, intuition, emotions, passion, desires, pleasure, aesthetics, joy, values, morality, experience, wisdom, and judgement. All of these attributes are essential elements of or conditions for human intelligence, in our opinion. In a certain sense, they are the ground in which human intelligence operates. Stripped of these qualities as is the case with AI all that is left of intelligence is logic, a figure without a ground according to McLuhan as we have already mentioned. If those that desire to create a human level of intelligence in a machine they will have to find a way to duplicate the above list of characteristics that we believe define human intelligence.

To the long list above of human characteristics that we have suggested contributes to human intelligence we would also add humor based on the following report of the social interaction of Stan Ulam and John von Neumann, the very first scholars to entertain the notion of the singularity.

> Von Neumann's closest friend in the United States was mathematician Stanislaw Ulam. A later friend of Ulam's, Gian-Carlo writes: "They would spend hours on end gossiping and giggling, swapping Jewish jokes, and drifting in and out of mathematical talk". When von Neumann was dying in hospital, every time Ulam would visit he would come prepared with a new collection of jokes to cheer up his friend (https://en.wikipedia.org/wiki/John_von_Neumann).

Humor entails thinking out of the box, a key ingredient of human intelligence. Humor specifically works by connecting elements that are not usually connected, as is also the case with creative thinking. All of the super intelligent people we have known invariably have a great sense of humor. Who can doubt the intelligence of the comics Robin Williams and Woody Allen, or the sense of humor of physicists Albert Einstein and Richard Feynman?

There are computers that can calculate better than us, and in the case of IBM's Big Blue, play chess better than us, but Big Blue is a one-trick pony that is incapable of many of the facets of thinking that we regard as essential for considering someone intelligent. Other examples of computers that exceeded humans in game playing are Google's AlphaGo beating the human Go champion and IBM's Watson beating the TV Jeopardy champion. In the case of Watson, it won the contest but it had no idea of what the correct answers it gave meant and it did not realize that it won the contest, nor did it celebrate its victory. What kind of intelligence is that? A very specialized and narrow kind for sure.

Perhaps the biggest challenge to our skepticism vis-à-vis the Singularity is a recent feat by the non-profit organization OpenAI with the mission of openly sharing its AI research. They developed an AI machine that can play games against itself and thereby find the optimum strategy for winning the game. It played GO against itself for three days and when it was finished, it was able to beat the original AlphaGo computer that had beat the human Go champion. In fact, it played 100 matches against AlphaGo and it won them all. AI devices that can beat humans at ruled base games parallels the fact that computers can calculate far faster and far better than any human. The other aspect of computers beating humans playing games is that a game is a closed system, whereas life and reality is an open system [28].

Intelligence, at least the kind that we value, involves more than rule based activities and is not limited to closed systems, but operates in open systems. All of the breakthroughs in science and the humanities involve breaking the rules of the previous paradigms in those fields. Einstein's theory of relativity and quantum theory did not follow the rules of classical physics. As for the fine arts, there are no rules. Both the arts and the sciences are open systems.

The idea that a computer can have a level of imagination or wisdom or intuition greater than humans can only be imagined, in our opinion, by someone who is unable to understand the nature of human intelligence. It is not our intention to insult those that have embraced the notion of the technological singularity, but we believe that this fantasy is dangerous and has the potential to mislead the developers of computer technology by setting up a goal that can never be reached, as well as devalue what is truly unique about the human spirit and human intelligence.

It is only if we lower our standards as to what constitutes human intelligence, will computers overtake their human creators as advocates of AGI and the technological singularity suggest. Haim Harari [29] p. 434 put it very succinctly when he wrote that he was not worried about the world being controlled by thinking machines but rather he was "more concerned about a world led by people who think like machines, a major emerging trend of our digital society". In a similar vein, Devlin [30] p. 76 claims that computers cannot think, they can only make decisions, and that he further claims, is the danger of AGI, namely, decisions that are made without thought.

4. The 3.5 Billion Year Evolution of Human Intelligence

Many of the shortcomings of AGI as compared to human intelligence is due to the fact that human beings are not just logic machines, but they are flesh and blood organisms that perceive their environment, have emotions, goals, and have the will to live. These capabilities took 3.5 billion years of evolution to create.

Kate Jeffrey [31] pp. 366–369 suggests that it would be "an immense act of hubris" to achieve the same level of human intelligence in a machine. She asks, "Can we do better than 3.5 billion years of evolution did with us?"

Anthony Aguirre [32] pp. 212–214 remarks that, "human minds are incredibly complex but have been battle tested into (relative) stability over eons of evolution in a variety of extremely challenging environments. AGI computers, on the other hand, cannot be built in the millions or billions, and how many generations of them need to be developed before they achieve the stability of the human brain".

As S. Abbas Raza [33] pp. 257–259 asks, can any process other than Darwinian evolution produce "teleological autonomy akin to our own".

Gordon E. Moore [34], cofounder of Intel Corp, cofounder of Fairchild Semiconductor and the Moore of Moore's law is also a skeptic.

> I don't believe this kind of thing is likely to happen, at least for a long time. And I don't know why I feel that way. The development of humans, what evolution has come up with, involves a lot more than just the intellectual capability. You can manipulate your fingers and other parts of your body. I don't see how machines are going to overcome that overall

> gap, to reach that level of complexity, even if we get them so they're intellectually more capable than humans.

4.1. Human Intelligence and the Figure/Ground Relationship

In the Introduction, we indicated that human intelligence for us is not just a matter of logic and rationality, but that it also entails explicitly the following characteristics that we will now show are essential to human thought: purpose, objectives, goals, telos, caring, intuition, imagination, humor, emotions, passion, desires, pleasure, aesthetics, joy, curiosity, values, morality, experience, wisdom, and judgement. We will now proceed through this list of human characteristics and show how each is an essential component of human intelligence that would be difficult if not impossible to duplicate with a computer. These characteristics arise directly or indirectly because of the fact that humans have a sense of self that motivates these characteristics. Without a sense of self, who is it that has purpose, objectives, goals, telos, caring, intuition, imagination, humor, emotions, passion, desires, pleasure, aesthetics, joy, curiosity, values, morality, experience, wisdom, and judgement. How could a machine have any of these characteristics?

4.2. Human Thinking Is Not Just Logical and Rational, but It Is Also Intuitive and Imaginative

Just as some advocates of the Singularity look at figures without considering the ground in which they operate, they also do not take into account that human thought is not just logical and rational but it is also intuitive, imaginative and even sometimes irrational.

The earliest and harshest critic of AGI was Hubert Dreyfus who wrote the following series of books beginning in 1965: *Alchemy and AI* [4], *What Computers Can't Do* [5], *Mind over Machine: The Power of Human Intuition and Expertise in the Era of the Computer* [6], and *What Computers* Still *Can't Do* [7].

Dreyfus made a distinction between knowing-that which is symbolic and knowing-how, which is intuitive like facial recognition. Knowing-how depends on context, which he claimed is not stored symbolically. He contended that AI would never be able to capture the human ability to understand context, situation, or purpose in the form of rules. The reason being that human intelligence and expertise depended primarily on unconscious instincts rather than conscious symbolic manipulation. He argued that these unconscious skills would never be captured in formal rules that a computer could duplicate. Dreyfus's critique parallels McLuhan's notion of figure/ground analysis. Just as McLuhan claimed that the ground was subliminal, Dreyfus also claims that the ground of human thought are unconscious instincts that allow us to instantly and directly arrive at a thought without going through a conscious series of logical steps or symbolic manipulations. Another way of formulating Dreyfus's insight is in terms of emergence theory. The human mind and its thought processes are emergent, non-reductionist phenomena. Computers, on the other hand, operate making use of a reductionist program of symbolic manipulations. AGI is linear and sequential, whereas human thinking processes are simultaneous and non-sequential. In McLuhan's terminology, computers operate in one thing at a time, visual space and the human mind operates in the simultaneity of acoustic space. Computers operate as closed systems, no matter how large the databases can be. Biotic systems, such as human beings, are created by and operate within an open system.

Atran [35] pp. 220–222 reminds us that computers "process information in ways opposite to humans' in domains associated with human creativity". He points out that Newton and Einstein imagined "ideal worlds without precedent in any past or plausible future experience ... Such thoughts require levels of abstraction and idealization that disregard, rather than assimilate" what is known. Dyson [36] pp. 255–256 strikes a similar note: "genuinely creative intuitive thinking requires non-deterministic machines that can make mistakes, abandon logic from one moment to the next and learn. Thinking is not as logical as we think".

Imagination and curiosity are uniquely human and are not mechanical one step at a time. Mechanically trying to program them as a series of logical steps is doomed to fail, since imagination and curiosity defy logic and a computer is bound by logic. Logic can inhibit creativity, imagination and curiosity. An example of how deviation from strictly logical thinking leads to new ideas is the story of the invention of zero. Parmenides argued that nothing changes because 'non-being' cannot be because it is a contradiction. As a result, nothing changes because if A changes into B, the A will not-be but since non-being cannot be nothing changes.

His use of logic led to a non-intuitive result but it impacted much of Ancient Greek thinking with all subsequent Greek philosophers adding something to their model of the world that was unchanging like Plato's forms and Aristotle's domain of the heavens. We believe that Parmenides argument that non-being could not be also explains why the Greeks, who were great at geometry, never came up with the idea of zero. Zero was an invention of the Hindu mathematicians who were not always very logical but very practical. When they recorded their abacus calculations, they used a symbol they called sunya (leave a space) to record 302 as 3 sunya 2, i.e., 3 in the 100 column, nothing in the ten column and 2 in the one column. They denoted sunya with a dot and later with a circle. Sunya allowed them to invent the place number system, which we call Arabic numbers. The Arabs adopted sunya and translated 'leave a space' into sifr. When the Italians borrowed sifr from the Arabs, they called it zefiro and later shortened it to zero. Zero, place numbers, negative numbers, and algebra all emerged from what Parmenides and his Greek compatriots would have called a logical error.

Lawrence Krauss, a physicist, expressed a wish for AGI machines: "I'm interested in what machines will focus on when they get to choose the questions as well as the answers" [37]. We are quite skeptical of this hope of Krauss given the following remark of Einstein and Infeld [38]: "The mere formulation of a problem is far more often essential than its solution, which may be merely a matter of mathematical or experimental skill. To raise new questions, new possibilities, to regard old problems from a new angle requires creative imagination and marks real advances in science". It is hard to imagine how a system of logic could develop a curiosity, as all logic can do is equate equivalent statements.

Kevin Kelly [39] also sees AGI machines tackling scientific questions: "To solve the current grand mysteries of quantum gravity, dark energy, dark matter, we'll probably need intelligences other than human". Kelly is not taking into account the imagination that is required to create a new paradigm. A new paradigm does arise from making a calculation, but from using one's imagination. If he had written his piece in 1904, he might have suggested that we needed AI to understand the result of the Michaelson-Morley experiment, which showed that there is no aether and the speed of light is the same in all frames of reference. Einstein did not use computation or logic to come up with the theory of relativity, only imagination. Einstein [40] remarked, "Imagination is more important than knowledge. For knowledge is limited to all we now know and understand, while imagination embraces the entire world, and all there ever will be to know and understand".

Gordon Kane [41] p. 219 reminds us that AI might be able to analyze data but scientists will still have to build technological devices to gather data, without which there is no science.

Rafaeli [42] pp. 342–344 also sees AGI computers someday making scientific progress: "Thinking needs data, information and knowledge but also requires communication and interaction. Thinking is about asking questions, not just answering them . . . For a machine to think it will need to be curious, creative and communicative". Rafaeli believes that machines will one day be communicative, but we wonder what will motivate them to do so or to be creative or curious for that matter.

4.3. Purpose, Objectives, Goals, Telos and Caring

> *Everything—a horse, a vine—is created for some duty... For what task, then, were you yourself created? A man's true delight is to do the things he was made for* —Marcus Aurelius

The set of five interrelated characteristics of purpose, objectives, goals, telos, and caring, define what it is to be a living organism. All living organisms and only this class of objects have these properties because they are the only entities that act in their own self-interest. One can define a living organism as an entity that has purpose, objectives, goals, and a telos, a will to live and to reproduce, an end. Telos, associated with Aristotle's fourth cause, is the purpose, goal, or objective that gives rise to an efficient cause that makes something happen. Computers do not have a will to live, a purpose, a goal, or objectives nor do they care about anything. They just function as they were designed to perform and as they are programmed by their human manufacturers and users. They do not reproduce like all forms of life, including bacteria and viruses. Because of their will to live, all organisms are caring; they care about finding nourishment and whether they live or die. Even single cell organisms like bacteria and more complex eukaryote microbes will communicate with each other when there is not an ample supply of nourishment to sustain themselves and form a slime medium that allows them to migrate to where there is more food. Bacteria also form slime to protect themselves from ingestion or desiccation. Slime molds, which are eukaryotes, also form cooperative slime colonies for finding nourishment and for reproduction. As more complex multi-cell organisms emerged more complex forms of caring emerged. Computers, on the other hand, have no capacity for caring. Caring, an emotional state, as we will discover below is key for creative thinking.

4.4. Intuition

Intuition is the clear conception of the whole at once —Johann Kaspar Lavater

Living organisms basically make decisions based on their intuition and not on a linear deductive use of logic or any other process for that matter. One exception are humans, who because of their capacity with symbolic language, sometimes make decisions through a process of reasoning, but for the most part their day to day decisions regarding their safety, their nourishment, their movement, their breathing, are all done intuitively. It is only when they are planning, building something, or solving a problem that they make use of logical reasoning. But intuition kicks in again when they are engaged in the following activities: solving a wicked problem; creating or composing music; making an art object like a painting or a piece of sculpture; performing in a play; dancing; engaged in sports; driving a car; flying a plane; or, sailing a boat. In most of these cases there would not be enough time to proceed through a chain of logical steps to make decisions upon. In the case of wicked problem solving, the solution lies in making assumptions that have never been made before.

Logic has nothing to do with making the assumptions upon which a chain of logical thinking is executed. Logic only helps one develop a solution based on the assumptions one has made. Imagining new assumption is an intuitive act not an act of reason or rationality. Thomas Kuhn has made a study of how new scientific breakthroughs are made. They are always made by someone new to the field, usually a young scientist who intuits the new paradigm by making an assumption that contradicts the logic of the old paradigm. This is why an AI device cannot solve a wicked problem because it operates as a logic closed system, and thus cannot intuit a new paradigm or a new set of assumptions. That requires a creative artistic-like or improvisation approach to science or problem solving. Improvisation cannot be achieved using logic. Logic constrains one's thinking, making improvisation and imagination impossible. Creative thinking is not rational and most times defies logic. Improvisation is about breaking the rules. Computers with their logical step by step processes cannot leap directly to a solution as is the case with an intuitive thinker.

The moment we, the authors, learned about the idea of a Singularity, we immediately sensed that it was wrong. As to those that hold a belief in the Singularity, it was also a result of their intuitive thinking. The explanation as to why the intuition of human agents can be so different is because they have different emotional needs. It is no accident that many Singularitarians are computer scientists, especially AI specialists.

4.5. Imagination

Imagination is an important component of intelligence —Albert Einstein

Imagination entails the creation of new images, concepts, experiences, or sensations in the mind's eye that have never been seen, perceived, experienced, or sensed through the senses of sight, hearing, and/or touch. Computers do not see, perceive, experience, or sense, as do humans, and therefore cannot have imagination. They are constrained by logic and logic is without images. Another way of describing imagination is to say it represents thinking outside the box. Well, is the box not all that we know and the equivalent ways of representing that knowledge using logic? Logic is merely a set of rules that allows one to show that one set of statements is equivalent to another. One cannot generate new knowledge using logic; one can only find new ways of representing it. Creativity requires imagination and imagination requires creativity and both creativity and imagination are intuitive, so once again we run up against another barrier that prevent computers from generating general intelligence.

Imagination is essential in science for creating a hypothesis to explain observed phenomena, and this part of the process of scientific thinking requires imagination, which is quite independent of logic. Logic comes into play when one used logic to determine the consequences of one's hypotheses that can be tested empirically. Devising ways to test one's hypotheses requires another, but quite different, kind of imagination.

Imagination is also a key element of artistic creation. The artist creates sensations for his or her audience that they (the audience) would not ordinarily experience.

4.6. Humor

> *He laughed to free himself from his minds bondage* —James Joyce

Humor is not so much a pre-requisite for intelligence as it is an indication of intelligence. To create or appreciate humor, one requires an imagination to see alternatives to one's expectations. The "incongruity theory (of humor) suggests that humor arises when logic and familiarity are replaced by things that don't normally go together [43]". Given that a computer would not recognize or even create this kind of incongruity, then they would not only lack a sense of humor, but would not be able to assemble such incongruities, which are an essential part of imagination, and hence, intelligence.

4.7. Emotions

> *The fairest thing we can experience is the mysterious. It is the fundamental emotion which stands at the cradle of true art and true science* —Albert Einstein

Humans experience a wide variety of emotions, some of which, as Einstein suggests, motivate art and science. Emotions, which are a psychophysical phenomenon, are closely associated with pleasure (or displeasure); passion; desires; motivation; aesthetics; and, and joy. Every human experience is actually emotional. It is a response of the body and the brain. Every experience is about what action to take. Acting to do it again or not do it again (private communication Terry Deacon).

Emotions play an essential part in human thinking as neuroscientist Antonio Damasio has shown:

> Damasio's studies showed that emotions take (or play) an important part in the human rational thinking mechanism [44] p. 326.

> For decades, biologists spurned emotion and feeling as uninteresting. But Antonio Damasio demonstrated that they were central to the life-regulating processes of almost all living creatures. Damasio's essential insight is that feelings are "mental experiences of body states", which arise as the brain interprets emotions, themselves physical states arising from the body's responses to external stimuli. (The order of such events is: I am threatened, experience fear, and feel horror.) He has suggested that consciousness, whether the primitive "core consciousness" of animals or the "extended" self-conception of humans, requiring autobiographical memory, emerges from emotions and feelings [45].

Terrence Deacon [9] pp. 512, 533 in *Incomplete Nature* also claims that emotions are essential for mental activities:

> Emotion . . . is not merely confined to such highly excited states as fear, rage, sexual arousal, love, craving, and so forth. It is present in every experience, even if often highly attenuated,

> because it is the expression of the necessary dynamic infrastructure of all mental activity . . . Emotion . . . is not some special feature of brain function that is opposed to cognition

Computers are incapable of emotions, which in humans, are inextricably linked to pleasure and pain because they have no pain nor any pleasure, and hence there is nothing to get emotional about. In addition, they have none of the chemical neurotransmitters, which is another reason why computers are incapable of emotions and the drives that are associated with them. Without emotions, computers lack the drive that are an essential part of intelligence and the striving to achieve a purpose, an objective or a goal. Emotions play a key role in curiosity, creativity, aesthetics, which are three other factors that are essential for human intelligence.

Singularitarians are essentially dualists that embrace the dualisms between body and mind and between reason and emotion. They are the last of the behaviorists that have replaced the Skinner box with a silicon box (today's computers). The mind is not just the brain, and the brain is not just a network of neurons operating as logic gates. The human mind extends into the body, is extended into our language according to Logan [46], and extended into our tools according to Clark and Chalmers [47].

4.8. Curiosity

> *Curiosity is one of the permanent and certain characteristics of a vigorous intellect* —Samuel Johnson

> *I have no special talent. I am only passionately curious*. —Albert Einstein

Curiosity is both an emotion and a behavior. Without the emotion of curiosity, the behavior of curiosity is not possible, and given that computers are not capable of emotions, then they cannot be curious, and hence lack an essential ingredient for intelligence. Curiosity entails the anticipation of reward, which in the brain comes in the form of neurotransmitters like dopamine and serotonin. No such mechanism exists in computers, and hence they totally lack native curiosity. Curiosity, if it exists at all, would have to be programmed into them. In fact, that is exactly what NASA did when it sent its Mars rover, aptly named Curiosity, to explore the surface of Mars.

Curiosity and intelligence are highly correlated. Advances in knowledge have always been the result of someone's curiosity. Curiosity is a characteristic that only a living organism can possess and no living organism is more curious than humans. How could a computer create new forms of knowledge without being curious? But that level of curiosity would have to the curiosity of the programmers who create the AGI creature. Since the curiosity programmed into the AGI device cannot exceed native human curiosity, this represents a real barrier to the achievement of the Singularity.

4.9. Creativity and Aesthetics

The role of creativity and aesthetics in the fine arts is rather obvious but it also plays a key role in science, engineering, product design and general problem solving. Humans solve problems and make discoveries using both out of the box and elegant thinking that defy the logical, one thing at a time, line of thought characteristic of uncreative thinkers and computers. Creativity is a passionate emotion-filled pursuit in which the creator cares about their creation whether it be a practical well-designed product, a scientific theory or an objet d'art.

Terence Deacon [9], pp. 91–92 also weighs in on the question of computers and creativity, which he shows if it were the case that a series of logical steps is all that is required to be creative then we live in a pre-determined world in which there is no free will, which is not how we find this world:

> Consider, however, that to the extent that we map physical processes onto logic, mathematics, and machine operation, the world is being modeled as though it is preformed, with every outcome implied in the initial state. But as we just noted, even Turing recognized that this mapping between computing and the world was not symmetrical. Gregory Bateson [48] p. 58 (1972, 58) explains this well:

> > In a computer, which works by cause and effect, with one transistor triggering another, the sequences of cause and effect are used to simulate logic. Thirty years ago, we used to ask: Can a computer simulate all the processes of logic? The answer was "yes", but the question was surely wrong. We should have asked: Can logic simulate all sequences of cause and effect? The answer would have been: "no".

> When extrapolated to the physical world in general, this abstract parallelism has some unsettling implications. It suggests notions of predestination and fate: the vision of a timeless, crystalline, four-dimensional world that includes no surprises. This figures into problems of explaining intentional relationships such as purposiveness, aboutness, and consciousness, because as theologians and philosophers have pointed out for centuries, it denies all spontaneity, all agency, all creativity, and makes every event a passive necessity already prefigured in prior conditions. It leads inexorably to a sort of universal preformationism.

Aesthetic also plays a role and not just in the fine arts and design but also in science and engineering. Einstein once remarked, "the only physical theories that we are willing to accept are the beautiful ones". Herman Bondi [49] confirmed this attitude of Einstein's when he wrote,

> What I remember most clearly was that when I put down a suggestion that seemed to me cogent and reasonable, Einstein did not in the least contest this, but he only said, 'Oh, how ugly.' As soon as an equation seemed to him to be ugly, he really rather lost interest in it and could not understand why somebody else was willing to spend much time on it. He was quite convinced that beauty was a guiding principle in the search for important results in theoretical physics.

The idea of a computer having a sense of aesthetics is preposterous given that the feeling of beauty is an emotion and computers cannot have emotions. Emotions involve the nervous system and the psyche, and since the computers do not have a nervous system or a psyche, they cannot feel emotions like the emotion of beauty. So, once again, it is hard to imagine AGI competing with or developing anything like human intelligence.

4.10. Values and Morality

Because a computer has no purpose, objectives, or goals, it cannot have any values as values are related to one's purpose, objectives, and goals. As is the case with curiosity, values will have to be programmed into a computer, and hence the morality of the AGI device will be determined by the values that are programmed into it, and hence the morality of the AGI device will be that of its programmers. This gives rise to a conundrum. Whose values will be inputted, and who will make this decision, a critical issue in a democratic society. Not only that, but there is a potential danger. What if a terrorist group or a rogue state were to create or gain control of a super-intelligent computer or robot that could be weaponized. Those doing AGI research cannot take comfort in the notion that they will not divulge their secrets to irresponsible parties. Those that built the first atomic bomb thought that they could keep their technology secret, but the proliferation of nuclear weapons of mass destruction became a reality. When considering how many hackers are operating today, is not the threat of super-intelligent AGI agents a real concern?

Intelligence, artificial or natural, entails making decisions, and making decisions requires having a set of values. So, once again, as was the case with curiosity, the decision-making ability of an AGI device cannot exceed that of human decision making as it will be the values that are programmed into the machine that will ultimately decide which course of action to take and which decisions are made.

5. Artificial Intelligence and the Figure/Ground Relationship

Another insight of McLuhan's, namely the relationship of figure and ground, can help to explain why so many intelligent scientists can go so far astray in advocating the idea that a machine can think. The meaning of any "figure" according to McLuhan, whether it is a technology, an institution, a communication event, a text, or a body of ideas, cannot be determined if that figure is considered in isolation from the ground or environment in which it operates. The ground provides the context from which the full meaning or significance of a figure emerges. The following examples illustrate the way in which the context can transform the meaning of a figure: a smokestack belching smoke, once a symbol of industrial progress, is today a symbol of pollution; the suntan, once a sign of hard work in the fields, is now a symbol of affluence and holidaying and will probably evolve into a symbol of reckless disregard for health and the risk of skin cancer.

We believe that a computer operates strictly on the figure of the problem that it is asked to solve. It has no idea of the ground or the context in which the problem arose. It is therefore not thinking, but merely manipulating symbols for concepts that it has never experienced and for which it has not had a perceptual experience, hence no emotions, no caring. Basically, the machine does not give a damn. Thinking entails making use of a bit of wisdom, which only can be acquired with experience that has both an intellectual and an emotional component, the latter of which is impossible for a machine. In other words, the machine cannot have emotions, feel love, pain, or regret, or have a sense of what is just or beautiful and hence can never become wise. The computer can only manipulate the figure of a problem, and it really has no clue about the ground or the context of the problem.

Can machines think? Just because a machine can calculate or compute faster than a human does not mean that it is thinking. It is just carrying out computations that a human who programs the computer has asked it to do. Before computers humans used abacuses and slide rules to facilitate their calculations. It never occurred to anyone to suggest that the abacuses or slide rules could think. No—they only carried out operations that their human operators made then perform. According to Andy Clark [50], these devices became extensions of the human mind. Computers, like abacuses and slide rules, only carry out operations their human operators/programmers ask them to do, and as such, they are extensions of the minds of their operators/programmers. They cannot think as they have no free will, in fact they have no will at all.

We would imagine that proponents of AGI or strong AI would claim that free will is an illusion and that our argument is simply a category error. Well, if there is no such thing as free will, then there is no difference between a human and a computer, as both are subject to the laws of physics. If that is the case, why do we value human life more than computer life. Should a person be charged with murder who destroys a computer as we have done when we abandoned our former out of date computers to recycling. The answer is, of course, no, but it does raise the question, would an AGI computer in the post-Singularity days be protected against murder like the humans that they will replace. Which raises another question: If post-Singularity computers control society how will they enforce the law.

Our position that computers or machines are mindless is supported by many of our confreres who are also Singularity skeptics. Here is a sample of a few thinkers who believe that AI computers are mindless:

> Much of the power of artificial intelligence stems from its mindlessness... Unable to question their own actions or appreciate the consequences of their programming—unable to understand the context in which they operate—they can wreak havoc either as a consequence of flaws in their programming or through the deliberate aims of their programmers [51] p. 59.

> Silicon-based information processing requires interpretation by humans to become meaningful and will for the foreseeable future. We have little to fear from thinking machines and more to fear from the increasingly unthinking humans who use them [52] p. 89.

> Machines (at least so far, and I don't think this will change with a Singularity) lack vision [53] p. 93.
>
> Machines (humanly constructed artifacts) cannot think, because no machine has a point of view—that is, a unique perspective on the worldly referents of its internal symbolic logic [54] p. 7.
>
> Machines don't think. They approximate functions. They turn inputs into outputs ... much of machine thinking is just machine hill climbing ... Tomorrow's machines will look a lot like today's—old algorithms running on faster machines [55] pp. 423–426.

Basically, as each of the five critics above have pointed out AGI is a figure without a ground and a figure without a ground is dangerous because it lacks meaning because of a lack of context as Marshall McLuhan has observed.

In each of these quotes the ground that is missing to support the machines mechanical processing is: for Carr [51] "appreciate the consequences of their programming"; for Fitch [52] "interpretation"; for Pepperberg "vision"; for Trehub "a point of view" and for Kosko thinking substituted by "machine hill climbing".

5.1. The Turing Test

Alan Turing, in 1950, developed criteria to determine if a machine could exhibit intelligent behavior. The Turing test, as it became to be known, was whether a human dialoging with a computer in a text only channel could determine whether or not they were in conversation with a machine or a human. If the human could not tell whether it was a human or a computer, the computer passed the Turing test. The Turing test might be a necessary condition for a computer possessing human-like intelligence, but it certainly is not a sufficient condition. A smart interlocutor, however, could easily determine whether they were conversing with a machine or a human by asking the following personal questions: Have you ever been in love?; Who are you closest to in your family?; What is your gender?; What gives you joy and why?; What are your goals in life?; What is the ethnic origin of your family?; What sports do you enjoy participating in and why? What was the happiest moment in your life and the saddest? The Turing test is not really a test of intelligence, it is a test of whether a programmer can fool a human to believe it is also a human. In other words, it is a magician's trick (private communication with Terry Deacon).

5.2. Machines Do Not Network and They Have No Theory of Mind

Stanislaw Dehaene [56] pp. 223–225 points out that machines lack two essential functions for intelligent thinking a global workspace and a theory of mind. The human mind knows what it knows. Although it has specialized modules where the information is accessible by all of the brain, similar to Pribram's holographic brain with its holographic storage network. Computer modules do not have holographic access to information, and a computer module, unlike a human brain module, is not aware of the information in the other modules.

The human mind is capable of responding and attending to other humans, but AI does not have this capability to respond and attend to its users. Computers operate as figure without ground. Ridley [57] pp. 226–227 points out "human intelligence is not individual thinking at all but collective, collaborative and distributive intelligence". Networking with humans is possible because humans have language which animals and computers do not have. Control of fire and the living in communal groups led to verbal language and networking. Ridley claims the only hope for the AGI will be individual AGI configured computers that are interlinked by the Internet.

Shafir [58] pp. 300–301 points out that since an AGI computer cannot have a Theory of Mind, it will never be able to achieve the level of human intelligence. A theory of mind emerges when a human realizes other humans think the way they do. Since a computer cannot think like a human, it will never be able to develop a theory of mind.

5.3. AI Have No Goals, Feelings or Emotions and Hence Cannot Act and They Do Not Care

A computer cannot experience pain, pleasure, or joy, and therefore has no motivation, no goals, no desire to communicate according to Enfield [59] pp. 3917–3998: "Machines comply, but they don't cooperate" because they have no goals.

David Gelernter [60] pp. 80–83. "Philosophers and scientists wish that the mind was nothing but thinking and that *feeling* or *being* played no part. They wished so hard for it to be true that they finally decided it was. Philosophers are only human".

Edward Slinerland [61] pp. 345–346 regards AGI computers as "directionless intelligences because AI systems are tools not organisms ... No matter how good they become [doing things] they don't actually want to do any of these things ... AI systems in and of themselves are entirely devoid of intentions or goals ... Motivational direction is the product of natural selection working on biological organisms".

Since computers operate using an either/or, 0 or 1, true or false logic, they are not capable of metaphor, which Stuart Kauffman [62] pp. 507–509 claims is the basis of human creative thought, whether mathematical, artistic or scientific.

Abigail Marsh [63] pp. 415–417 cites a patient that is incapable of any emotions because of damage to his ventro-medial prefrontal cortex. The patient was unable to make use of his intelligence and knowledge residing in other unaffected areas of his brain because without his emotional capacity he was "unable to make decisions or take action". No emotions—no actions. One, therefore, cannot expect any thought or initiative from an AGI device. It is apathetic, has no capacity for emotions and has no initiative unless instructed by a human. It all comes down to the fact that there is no reward for a computer, and hence no motivation. Therefore, the notion that is expressed by some advocates of the Singularity that AGI computers could take over the human race is without basis. Roy Baumeister [64] p. 73 comes to a similar conclusion.

Johnathan Gottschall [65] pp. 179–180 points out that the success of AGI computers to generate compelling stories has been a dismal failure. The creators of great stories and other forms of artistic expression are intelligent but they have another quality, which we will call soulfulness. By soul in this context we are expressing something that is not supernatural but something that has a strong emotional component to it, in addition to the analytic intelligence that an artist must also possess. Musicologists have formulated the rules that Mozart used in composing his music, and then fed those rules into a computer along with a simple Mozart melody. The result was music that sounded a bit like Mozart but had none of the emotional and aesthetic appeal of the music that Mozart actually composed. Mozart has soul and passion and the computer has a melody, a set of rules and the ability to combine them, but not the ability to create beauty. An artist knows when to break the rules, whereas the computer can only stick to the rules. There is a parallel with creative science. The scientists that breaks new ground breaks the rules of the former Kuhnian paradigm by combining intelligence with intuition and creative imagination.

Saffo [66] pp. 206–208 suggests one of the motivation to create AGI is that "we want someone to talk to". This suggestion raises the question: but if we have the possibility to talk to each other, why would we need an AGI computer to fill in this need. Saffo also suggests that, "we of course will attribute feelings and rights to AIs—and eventually they will demand it". How can machines without a sense of self or without a will to live be able to demand anything which begs the question how will they arrive at a notion of rights unless that is programmed into them. Finally, the last straw for us is Saffo's contention that sentience is universal and not limited to living things. "It's just a small step to speculate about what trees, rocks—or AIs—think". I guess that means trees and rocks have rights and we can talk to them. Maybe the tree huggers know something that we do not know.

Kosslyn [67] pp. 228–230 posits that AGI computers will want to have purpose, and as a consequence, they will want to support and elevate the human condition. This is another example of a Singularity advocate assuming that a machine without any capacity for emotions having the capacity to desire something that would give it pleasure. We contend that without emotions and without the ability to feel pleasure, there can be no desire and no purpose.

6. Decision Making, Experience, Judgement and Wisdom

The only Source of Knowledge is Experience. —Albert Einstein

If the challenges of programming an AGI device with a set of values and a moral compass that represents the will of the democratic majority of society is achieved, there is still the challenge of whether the AGI device still has the judgment and wisdom to make the correct decision. In other words, is it possible to program wisdom into a logical device that has no emotions and has no experiences upon which to base a decision. Wisdom is not a question of having the analytic skills to deal with a new situation but rather having a body of experiences to draw upon to guide one's decisions. How does one program experience into a logic machine?

Intelligence requires the ability to calculate or to compute, but the ability to calculate or compute does not necessarily provide the capability to make judgments and decisions unless values are available, which for an AGI device, requires input from a human programmer.

7. Conclusions: How Computers Will Make Us Humans Smarter

Douglas Rushkoff [12] pp. 354–355 invites us to consider computers not as figure but as ground. He suggests that the leap forward in intelligence will not be in AGI configured computers that have the potential to be smarter than us humans, but in the environment that computers create. Human intelligence will increase by allowing human minds to network and create something greater than what a single human mind can create, or what a small group of minds that are co-located can create. The medieval university made us humans smarter by bringing scholars in contact with each other. The city is another example of a medium that allowed thinkers and innovators to network, and hence increase human intelligence. The printing press had a similar impact. With networked computer technology, a mind with a global scale is emerging.

In the past, schools of thought emerged that represented the thinking of a group or team of scholars. They were named after cities. What is emerging now are schools of thought and teams of scholars that are not city-based but exist on a global scale. An example is that we once talked about the Toronto school of communication and media studies consisting of scholars, such as Harold Innis, Marshall McLuhan, Ted Carpenter, Eric Havelock, and Northrop Fry that lived in Toronto and communicated with each other about media and communications. A similar New York school of communication emerged with Chris Nystrom, Jim Carey, John Culkin, Neil Postman, and his students at NYU. Today, that tradition lives on but not as the Toronto School or the New York School, but as the Media Ecology School, with participants in every part of the world. This is what Rushkoff [12] was talking about in his article "The Figure or Ground", where he pointed out that it is the ground or environment that computers create, and not the figure of the computer by itself that will give rise to intelligence greater than a single human. He expressed this idea as follows: "Rather than towards machines that think, I believe we are migrating toward a networked environment in which thinking is no longer an individual activity nor bound by time and space".

Marcelo Gleiser [68] pp. 54–55 strikes a similar chord to that of Doug Rushkoff when he points out that many of our technologies act as extensions of who we are. He asks: "What if the future of intelligence is not outside but inside the human brain? I imagine a very different set of issues emerging from the prospect that we might become super-intelligent through the extension of our brainpower by digital technology and beyond—artificially enhanced human intelligence that amplifies the meaning of being human".

Author Contributions: A.B. and R.K.L. wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chalmers, D. The Singularity: A Philosophical Analysis. *J. Conscious. Stud.* **2010**, *17*, 7–65.
2. IEEE Spectrum. Special Report: The Singularity. 2008. Available online: http://spectrum.ieee.org/static/singularity (accessed on 15 January 2016).
3. Brockman, J. (Ed.) *What to Think About Machines that Think*; Harper Perennial: New York, NY, USA, 2015.
4. Dreyfus, H. *Alchemy and AI*; RAND Corporation: Santa Monica, CA, USA, 1965.
5. Dreyfus, H. *What Computers Can't Do*; MIT Press: New York, NY, USA, 1972.
6. Dreyfus, H. *What Computers Can't Do*, 2nd ed.; MIT Press: New York, NY, USA, 1979.
7. Dreyfus, H. *What Computers Still Can't Do*; MIT Press: New York, NY, USA, 1992.
8. Braga, A.; Logan, R. Communication, Information and Pragmatics. In *Encyclopedia of Information Science and Technology*; Khosrow-Pour, M., Ed.; IGI Global: Hershey, PA, USA, 2017.
9. Deacon, T. *Incomplete Nature: How Mind Emerged from Matter*; WW Norton and Company: New York, NY, USA, 2012.
10. McLuhan, M. *Understanding Media: Extensions of Man*; McGraw Hill: New York, NY, USA, 1964.
11. McLuhan, E. *Media and Formal Cause*; NeoPoiesis Press: New York, NY, USA, 2011.
12. Rushkoff, D. Figure or Ground? In *What to Think about Machines that Think*; Brockman, J., Ed.; Harper Perennial: New York, NY, USA, 2015; pp. 354–355.
13. Bratton, B. The Evolution Revolution. Available online: http://techonomy.com/wp-content/uploads/2016/12/The-Evolution-Revolution-with-Ray-Kurzweil-Benjamin-H-Bratton-and-Vivienne-Ming.pdf (accessed on 27 November 2017).
14. Kurzweil, R. *The Singularity Is Near*; Viking Books: New York, NY, USA, 2005.
15. Gibson, K. *Ethics and Business: An Introduction*; Cambridge University Press: Cambridge UK, 2007.
16. Sellen, A.; Harper, R. *The Myth of the Paperless Office*; MIT Press: Cambridge, MA, USA, 2003.
17. Pinker, S. Thinking Does Not Imply Subjugating. In *What to Think about Machines that Think*; Brockman, J., Ed.; Harper Perennial: New York, NY, USA, 2015; pp. 5–8.
18. Tipler, F. If You Can't Beat 'Em, Join 'Em. In *What to Think about Machines that Think*; Brockman, J., Ed.; Harper Perennial: New York, NY, USA, 2015; pp. 17–18.
19. Lisi, A.G. I, for One, Welcome our Machine Overlords. In *What to Think about Machines that Think*; Brockman, J., Ed.; Harper Perennial: New York, NY, USA, 2015; pp. 22–24.
20. McCorduck, P. An Epochal Human Event. In *What to Think about Machines that Think*; Brockman, J., Ed.; Harper Perennial: New York, NY, USA, 2015; pp. 51–53.
21. Harris, S. Can We Avoid a Digital Apocalypse? In *What to Think about Machines that Think*; Brockman, J., Ed.; Harper Perennial: New York, NY, USA, 2015; pp. 408–411.
22. Paul, G. What will AI's Think of Us. In *What to Think about Machines that Think*; Brockman, J., Ed.; Harper Perennial: New York, NY, USA, 2015; pp. 391–393.
23. Croak, J. Fear of a God, Redux. In *What to Think about Machines that Think*; Brockman, J., Ed.; Harper Perennial: New York, NY, USA, 2015; pp. 498–499.
24. Hofstadter, D. Tech Luminaries Address Singularity. 2008. Available online: http://spectrum.ieee.org/computing/hardware/tech-luminaries-address-singularity (accessed on 15 March 2016).
25. Horgan, J. The Consciousness Conundrum. 2008. Available online: http://spectrum.ieee.org/biomedical/imaging/the-consciousness-conundrum (accessed on 3 March 2016).
26. Hardy, Q. Interesting Thoughts brought to Mind. In *What to Think about Machines that Think*; Harper Perennial: New York, NY, USA, 2015; pp. 190–193.
27. Taylor, T. Denkraumverlust. In *What to Think about Machines that Think*; Brockman, J., Ed.; Harper Perennial: New York, NY, USA, 2015; pp. 251–254.
28. Quach, K. How DeepMind's AlphaGo Zero Learned All by Itself to Trash World Champ AI AlphaGo. Available online: https://www.theregister.co.uk/2017/10/18/deepminds_latest_alphago_software_doesnt_need_human_data_to_win (accessed on 18 October 2017).
29. Harari, H. Thinking about People Who Think Like Machines. In *What to Think About Machines that Think*; Brockman, J., Ed.; Harper Perennial: New York, NY, USA, 2015; p. 434.
30. Devlin, K. Leveraging Human Intelligence. In *What to Think About Machines that Think*; Brockman, J., Ed.; Harper Perennial: New York, NY, USA, 2015; pp. 74–76.

31. Jeffrey, K. In Our Image. In *What to Think about Machines that Think*; Brockman, J., Ed.; Harper Perennial: New York, NY, USA, 2015; pp. 366–369.
32. Aguirre, A. The Odds on AI. In *What to Think about Machines that Think*; Brockman, J., Ed.; Harper Perennial: New York, NY, USA, 2015; pp. 212–214 & 342–344.
33. Raza, S.A. The Value of Artificial Intelligence. In *What to Think about Machines that Think*; Brockman, J., Ed.; Harper Perennial: New York, NY, USA, 2015; pp. 257–259.
34. Moore, G. Tech Luminaries Address Singularity. 2008. Available online: http://spectrum.ieee.org/computing/hardware/tech-luminaries-address-singularity (accessed on 25 May 2016).
35. Atran, S. Are We Going in the Wrong Direction? In *What to Think about Machines that Think*; Brockman, J., Ed.; Harper Perennial: New York, NY, USA, 2015; pp. 220–222.
36. Dyson, G. Analog, the Revolution that Dares Not Speak Its Name. In *What to Think about Machines that Think*; Brockman, J., Ed.; Harper Perennial: New York, NY, USA, 2015; pp. 255–256.
37. Krauss, L.M. What Me Worry? In *What to Think about Machines that Think*; Brockman, J., Ed.; Harper Perennial: New York, NY, USA, 2015; pp. 171–174.
38. Einstein, A.; Infield, L. *The Evolution of Physics*; Simon and Schuster: New York, NY, USA, 1938.
39. Kelly, K. Call Them Artificial Aliens. In *What to Think about Machines that Think*; Brockman, J., Ed.; Harper Perennial: New York, NY, USA, 2015; pp. 245–247.
40. Einstein, A. What Life Means to Einstein: An Interview by George Sylvester Viereck. *The Saturday Evening Post*, 26 October 1929, p. 117.
41. Kane, G. We Need More than Thought. In *What to Think about Machines that Think*; Brockman, J., Ed.; Harper Perennial: New York, NY, USA, 2015; p. 219.
42. Rafaeli, S. The Moving Goalposts. In *What to Think about Machines that Think*; Brockman, J., Ed.; Harper Perennial: New York, NY, USA, 2015.
43. Ma, M. The Power of Humor in Ideation and Creativity. *Psychology Today*, 16 June 2014. Available online: www.psychologytoday.com/blog/the-tao-innovation/201406/the-power-humor-in-ideation-and-creativity (accessed on 30 July 2016).
44. Martınez-Miranda, J.; Aldea, A. Emotions in Human and Artificial Intelligence. *Comput. Hum. Behav.* **2005**, *21*, 323–341. [CrossRef]
45. Potin, J. The neuroscientist Antonio Damasio explains how minds emerge from emotions and feelings. *Technology Review*, 17 June 2014. Available online: https://www.technologyreview.com/s/528151/the-importance-of-feelings (accessed on 27 June 2015).
46. Logan, R.K. *The Extended Mind: The Origin of Language and Culture*; University of Toronto Press: Toronto, ON, Canada, 2007.
47. Clark, A.; Chalmers, D. The extended mind. *Analysis* **1998**, *58*, 10–23. [CrossRef]
48. Bateson, G. *Steps to An Ecology of Mind*; Random House/Ballantine: New York, NY, USA, 1972.
49. Bondi, H. *Einstein: The Man and His Achievement*; Withrow, G.J., Ed.; British Broadcasting Corporation: London, UK, 1973; p. 82.
50. Clark, A. *Natural-Born Cyborgs*; Oxford University Press: New York, NY, USA, 2003.
51. Carr, N. The Control Crisis. In *What to Think about Machines that Think*; Brockman, J., Ed.; Harper Perennial: New York, NY, USA, 2015; pp. 59–61.
52. Fitch, W. Tecumseh. Nano-Intentionality. In *What to Think about Machines that Think*; Brockman, J., Ed.; Harper Perennial: New York, NY, USA, 2015; pp. 89–92.
53. Trehub, A. Machines Cannot Think. In *What to Think about Machines that Think*; Harper Perennial: New York, NY, USA, 2015; p. 71.
54. Pepperberg, I. A Beautiful Visionary Mind. In *What to Think about Machines that Think*; Brockman, J., Ed.; Harper Perennial: New York, NY, USA, 2015; pp. 93–94.
55. Kosko, B. Thinking Machines—Old Algorithms on Faster Computers. In *What to Think about Machines that Think*; Brockman, J., Ed.; Harper Perennial: New York, NY, USA, 2015; pp. 423–426.
56. Dehaene, S. Two Cognitive Functions Machine Lack. In *What to Think about Machines that Think*; Brockman, J., Ed.; Harper Perennial: New York, NY, USA, 2015; pp. 223–225.
57. Ridley, M. Among the Machines Not Within the Machines. In *What to Think about Machines that Think*; Brockman, J., Ed.; Harper Perennial: New York, NY, USA, 2015; pp. 226–227.

58. Shafir, E. Blind to the Core of Human Experience. In *What to Think about Machines that Think*; Brockman, J., Ed.; Harper Perennial: New York, NY, USA, 2015; pp. 300–301.
59. Enfield, N.J. Machines Aren't into Relationships. In *What to Think about Machines that Think*; Brockman, J., Ed.; Harper Perennial: New York, NY, USA, 2015; pp. 97–98.
60. Gelernter, D. Why Can't 'Being' or 'Happiness' Be Computed. In *What to Think about Machines that Think*; Brockman, J., Ed.; Harper Perennial: New York, NY, USA, 2015; pp. 80–83.
61. Slinerland, E. Directionless Intelligence. In *What to Think about Machines that Think*; Brockman, J., Ed.; Harper Perennial: New York, NY, USA, 2015; pp. 345–346.
62. Kauffman, S. Machines that Think? Nuts! In *What to Think about Machines that Think*; Brockman, J., Ed.; Harper Perennial: New York, NY, USA, 2015; pp. 507–509.
63. Marsh, A. Caring Machines. In *What to Think about Machines that Think*; Brockman, J., Ed.; Harper Perennial: New York, NY, USA, 2015; pp. 415–417.
64. Baumeister, R. No 'I' and no Capacity for Malice. In *What to Think about Machines that Think*; Brockman, J., Ed.; Harper Perennial: New York, NY, USA, 2015; pp. 72–73.
65. Gottschall, J. The Rise of Storytelling Machines. In *What to Think about Machines that Think*; Brockman, J., Ed.; Harper Perennial: New York, NY, USA, 2015; pp. 179–180.
66. Saffo, P. What will the Place of Humans Be? In *What to Think about Machines that Think*; Brockman, J., Ed.; Harper Perennial: New York, NY, USA, 2015; pp. 206–208.
67. Kosslyn, S.M. Another Kind of Diversity. In *What to Think about Machines that Think*; Brockman, J., Ed.; Harper Perennial: New York, NY, USA, 2015; pp. 228–230.
68. Gleiser, M. Welcome to Your Transhuman Self. In *What to Think about Machines that Think*; Brockman, J., Ed.; Harper Perennial: New York, NY, USA, 2015; pp. 54–55.

information

Article

Countering Superintelligence Misinformation

Seth D. Baum

Global Catastrophic Risk Institute, PO Box 40364, Washington, DC 20016, USA; seth@gcrinstitute.org

Received: 9 September 2018; Accepted: 26 September 2018; Published: 30 September 2018

Abstract: Superintelligence is a potential type of future artificial intelligence (AI) that is significantly more intelligent than humans in all major respects. If built, superintelligence could be a transformative event, with potential consequences that are massively beneficial or catastrophic. Meanwhile, the prospect of superintelligence is the subject of major ongoing debate, which includes a significant amount of misinformation. Superintelligence misinformation is potentially dangerous, ultimately leading bad decisions by the would-be developers of superintelligence and those who influence them. This paper surveys strategies to counter superintelligence misinformation. Two types of strategies are examined: strategies to prevent the spread of superintelligence misinformation and strategies to correct it after it has spread. In general, misinformation can be difficult to correct, suggesting a high value of strategies to prevent it. This paper is the first extended study of superintelligence misinformation. It draws heavily on the study of misinformation in psychology, political science, and related fields, especially misinformation about global warming. The strategies proposed can be applied to lay public attention to superintelligence, AI education programs, and efforts to build expert consensus.

Keywords: artificial intelligence; superintelligence; misinformation

1. Introduction

At present, there is an active scholarly and public debate regarding the future prospect of artificial superintelligence (henceforth just *superintelligence*), which is artificial intelligence (AI) that is significantly more intelligent than humans in all major respects. While much of the issue remains unsettled, some specific arguments are clearly incorrect, and as such can qualify as *misinformation.* (As is elaborated below, arguments can qualify as misinformation even when the issues are unsettled.) More generally, misinformation can be defined as "false or inaccurate information" [1], or as "information that is initially presented as true but later found to be false" [2] (p. 1). This paper addresses the question of what can be done to reduce the spread of and belief in superintelligence misinformation.

While any misinformation is problematic, superintelligence misinformation is especially worrisome due to the high stakes involved. If built, superintelligence could have transformative consequences, which could be either massively beneficial or catastrophic. Catastrophe is more likely to come from a superintelligence built based on the wrong ideas—and it could also come from *not* building a superintelligence that would have been based on the *right* ideas, because a well-designed superintelligence could prevent other types of catastrophe, such that abstaining from building such a superintelligence could result in catastrophe. Thus, the very survival of the human species could depend on avoiding or rejecting superintelligence misinformation. Furthermore, the high stakes of superintelligence have the potential to motivate major efforts to attempt to build it or to prevent others from doing so. Such efforts could include massive investments or restrictive regulations on research and development (R&D), or plausibly even international conflict. It is important for these sorts of efforts to be based on the best available understanding of superintelligence.

Superintelligence is also an issue that attracts a substantial amount of misinformation. The abundance of misinformation may be due to the many high-profile portrayals of superintelligence in science fiction, the tendency for popular media to circulate casual comments about superintelligence made by various celebrities, and the relatively low profile of more careful scholarly analyses. Whatever the cause, experts and others often find themselves responding to some common misunderstandings [3–9].

There is also potential for superintelligence *disinformation*: misinformation with the intent to deceive. There is a decades-long history of private industry and anti-regulation ideologues promulgating falsehoods about socio-technological issues in order to avoid government regulations. This practice was pioneered by the tobacco industry in the 1950s and has since been adopted by other industries including fossil fuels and industrial chemicals [10,11]. AI is increasingly important for corporate profits and thus could be a new area of anti-regulatory disinformation [12]. The history of corporate disinformation and the massive amounts of profit potentially at stake suggest that superintelligence disinformation campaigns could be funded at a large scale and could be a major factor in the overall issue. Superintelligence disinformation could potentially come from other sources as well, such as governments or even concerned citizens seeking to steer superintelligence debates and practices in particular directions.

Finally, there is the subtler matter of the information that has not yet been established as misinformation, but is nonetheless incorrect. This misinformation is the subject of ongoing scholarly debates. Active superintelligence debates consider whether superintelligence will or will not be built, whether it will or will not be dangerous, and a number of other conflicting possibilities. Clearly, some of these positions are false and thus can qualify as misinformation. For example, claims that superintelligence will be built and that it will not be built cannot both be correct. However, it is not presently known which positions are false, and there is often no expert consensus on which positions are likely to be false. While the concept of misinformation is typically associated with information that is more obviously false, it nonetheless applies to these subtler cases, which can indeed be "information that is initially presented as true but later found to be false". Likewise, countering misinformation presents a similar challenge regardless of whether the misinformation is spread before or after expert consensus is reached (though, as discussed below, expert consensus can be an important factor).

In practical terms, the question then is what to do about it. There have been a number of attempts to reply to superintelligence misinformation in order to set the record straight [3–9]. However, to the best of the present author's knowledge, aside from a brief discussion in [12], there have been no efforts to examine the most effective ways of countering superintelligence misinformation. Given the potential importance of the matter, a more careful examination is warranted. That is the purpose of this paper. The paper's discussion is relevant to public debates about superintelligence, to AI education programs (e.g., in university computer science departments), and to efforts to build expert consensus about superintelligence.

In the absence of dedicated literature on superintelligence misinformation, this paper draws heavily on the more extensive research literature studying misinformation about other topics, especially global warming (e.g., [10,13,14]), as well as the general literature on misinformation in psychology, cognitive science, political science, sociology, and related fields (for reviews, see [2,15]). This paper synthesizes insights from these literatures and applies them to the particular circumstances of superintelligence. The paper is part of a broader effort to develop the social science of superintelligence by leveraging insights from other issues [12,16].

The paper is organized as follows. Section 2 presents some examples of superintelligence misinformation, in order to further motivate the overall discussion. Section 3 surveys the major actors and audiences (i.e., the senders and receivers) of superintelligence misinformation, in order to provide some strategic guidance. Section 4 presents several approaches for preventing the spread of superintelligence misinformation. Section 5 presents approaches for countering superintelligence misinformation that has already spread. Section 6 concludes.

2. Examples of Superintelligence Misinformation

It is often difficult to evaluate which information about superintelligence is false. This is because superintelligence is a possible future technology that may be substantially different from anything that currently exists, and because it is the subject of a relatively small amount of study. For comparison, other studies of misinformation have looked at such matters as whether Barack Obama was born in the United States, whether childhood vaccines cause autism, and whether Listerine prevents colds and sore throats [17]. In each of these cases, there is clear and compelling evidence pointing in one direction or the other (the evidence clearly indicates that Obama was born in the US, that vaccines do not cause autism, and that Listerine does not prevent colds or sore throats, despite many claims to the contrary in all three cases). Therefore, an extra degree of caution is warranted when considering whether a particular claim about superintelligence qualifies as misinformation.

That said, some statements about superintelligence are clearly false. For example, this statement from Steven Pinker: "As far as I know, there are no projects to build an AGI, not just because it would be commercially dubious, but also because the concept is barely coherent" [18]. The acronym AGI stands for artificial general intelligence, which is a form of AI closely associated with superintelligence. Essentially, AGI is AI that is capable of reasoning across a wide range of domains. AGI may be difficult to build, but the concept is very much coherent. Indeed, it has a substantial intellectual history and ongoing study [19], including a dedicated research journal (*Journal of Artificial General Intelligence*) and professional society (the Artificial General Intelligence Society). Furthermore, there are indeed projects to build AGI—one recent survey identifies 45, spread across many countries and institutions, including many for-profit corporations, the largest of which being DeepMind, acquired by Google in 2014 for £400 million, the Human Brain Project, an international project with $1 billion in funding from the European Commission, and OpenAI, a nonprofit with $1 billion in pledged funding [20]. (DeepMind and OpenAI explicitly identify as working on AGI. The Human Brain Project does not, but it is working on simulating the human brain, which is considered to be a subfield of AGI [19].) There is even an AGI project at Pinker's own university. (Pinker and the AGI project MicroPsi [21] are both at Harvard University.) Therefore, in the quoted statement, the "as far as I know" part may well be true, but the rest is clearly false. This particular point of misinformation is significant because it conveys the false impression that AGI (and superintelligence) is a nonissue, when in fact it is a very real and ongoing subject of R&D.

A more controversial matter is the debate on the importance of consciousness to superintelligence. Searle [22] argues that computers cannot be conscious and therefore, at least in a sense, cannot be intelligent, and likewise cannot have motivation to destroy humanity. Similar arguments have been made by Logan [23], for example. A counterargument is that the important part is not the consciousness a computer but its capacity to affect the world [4,24,25]. It has also been argued that AI could be harmful to humanity even if it is not specifically motivated to do so, because the AI could assess humanity as being in the way of it achieving some other goal [25,26]. The fact that AI has already shown the capacity to outperform humans in some domains is suggestive of the possibility for it to outperform humans in a wider range of domains, regardless of whether the AI is conscious. However, this is an ongoing area of debate, and indeed Chalmers [24] (p. 16) writes "I do not think the matter can be regarded as entirely settled". Regardless, there must be misinformation on one side or the other: computers either can be conscious or they cannot, and consciousness either matters for superintelligence or it does not. Additionally, many parties to the debate maintain that those who believe that consciousness or conscious motivation matter are misinformed [4,5,7–9], though it is not the purpose of this paper to referee this debate.

There are even subtler debates among experts who believe in the prospect of superintelligence. For example, Bostrom [25] worries that it would be difficult to test the safety of a superintelligence because it could trick its human safety testers into believing it is safe (the "treacherous turn"), while Goertzel [27] proposes that the safety testing for a superintelligence would not be so difficult because the AI could be tested before it becomes superintelligent (the "sordid stumble"; the term is from [28]). Essentially, Bostrom argues that an AI would become capable of deceiving humans before humans realize it is unsafe, whereas Goertzel argues the opposite. Only one of these views can be correct; the other would qualify as misinformation. More precisely, only one of these views can be correct for a given AI system—it is possible that some AI systems could execute a treacherous turn while others would make a sordid stumble. Which view is more plausible is a matter of ongoing study [28,29]. This debate is important because it factors significantly into the riskiness of attempting to build a superintelligence.

Many more additional examples could be presented, such as on the dimensionality of intelligence [3], the rate of progress in AI [7,8], the structure of AI goals [6–8], and the relationship between human and AI styles of thinking [6,8]. However, this is not the space for a detailed survey. Instead, the focus of this paper is on what to do about the misinformation. Likewise, this paper does not wish to take positions on open debates about superintelligence. Some positions may be more compelling, but arguments for or against them are tangential to this paper's aim of reducing the preponderance of misinformation. In other words, this paper strives to be largely neutral on which information about superintelligence happens to be true or false. The above remarks by Pinker will occasionally be used as an example of superintelligence misinformation because they are so clearly false, whereas the falsity of other claims is more ambiguous.

The above examples suggest two types of superintelligence misinformation: information that is already clearly false and information that may later be found to be false. In practice, there may be more of a continuum of how clearly true or false a piece of information is. Nonetheless, this distinction can be a useful construct for efforts to address superintelligence misinformation. The clearly false information can be addressed with the same techniques that are used for standard cases of misinformation, such as Obama's place of birth. The not-yet-resolved information requires more careful analysis, including basic research about superintelligence, but it can nonetheless leverage some insights from the misinformation literature.

The fact that superintelligence is full of not-yet-resolved information is important in its own right, and it has broader implications for superintelligence misinformation. Specifically, the extent of expert consensus is an important factor in the wider salience of misinformation. This matter is discussed in more detail below. Therefore, while this paper is mainly concerned with the type of misinformation that is clearly false, it will consider both types. With that in mind, the paper now starts to examine strategies for countering superintelligence misinformation.

3. Actors and Audiences

Some purveyors of superintelligence misinformation can be more consequential than others. Ditto for the audiences for superintelligence misinformation. This is important to bear in mind because it provides strategic direction to any efforts to counter the misinformation. Therefore, this section reviews who the important actors and audiences may be.

Among the most important are the R&D groups that may be building superintelligence. While they can be influential sources of ideas about superintelligence, they may be especially important as audiences. For example, if they are misinformed regarding the treacherous turn vs. the sordid stumble, then they could fail to correctly assess the riskiness of their AI system.

Also important are the institutions that support the R&D. At present, most AGI R&D groups are based in either for-profit corporations or universities, and some also receive government funding [20]. Regulatory bodies within these institutions could ensure that R&D projects are proceeding safely, such as via university research review boards [30,31]. Successful regulation depends on being well-informed about the nature of AGI and superintelligence and its prospects and risks. The same applies to R&D funding decisions by institutional funders, private donors, and others. Additionally, while governments are not presently major developers of AGI, except indirectly as funders, they could become important developers should they later decide to do so, and they meanwhile can play important roles in regulation and in facilitating discussion across R&D groups.

Corporations are of particular note due to their long history of spreading misinformation about their own technologies, in particular to convey the impression that the technologies are safer than they actually are [10]. These corporate actors often wield enormous resources and have a correspondingly large effect on the overall issue, either directly or by sponsoring industry-aligned think tanks, writers, and other intermediaries. At this time, there are only hints of such behavior by AI corporations, but the profitability of AI and other factors suggest the potential for much more [12].

Thought leaders on superintelligence are another significant group. In addition to the aforementioned groups, this also includes people working on other aspects of superintelligence, such as safety and policy issues, as well as people working on other (non-superintelligence) forms of AI, and public intellectuals and celebrities. These are all people who can have outsized influence when they comment on superintelligence. That influence can be on the broader public, as well as in quieter conversations with AGI/superintelligence R&D groups, would-be regulators, and other major decision-makers.

Finally, there is the lay public. The role of the public in superintelligence may be reduced due to the issue being driven by technology R&D that (for now at least) occurs primarily in the private sector. However, the public can play roles as citizens of governments that might regulate the R&D and as consumers of products of the corporations that host the R&D. The significance of the public for superintelligence is not well established at this time.

While the above groups are presented in approximate order of importance, it would not be appropriate to formally rank them. What matters is not the importance of the group but the quality of the opportunity that one has to reduce misinformation. This will tend to vary heavily by the circumstances of whoever is seeking to reduce the extent of superintelligence misinformation.

With that in mind, the paper now turns to strategies for reducing superintelligence misinformation.

4. Preventing Superintelligence Misinformation

The cliché "an ounce of prevention is worth a pound of cure" may well be an understatement for misinformation. An extensive empirical literature finds that once misinformation enters into someone's mind, it can be very difficult to remove.

Early experiments showed that people can even make use of information that they acknowledge to be false. In these experiments, people were told a story and then were explained that some information in the story is false. When asked, subjects would correctly acknowledge the information to be false, but they would also use it in retelling the story as if it were true. For example, the story could be a fire caused by volatile chemicals, and then it is later explained that there were no volatile chemicals present. Subjects would acknowledge that the volatile chemicals were absent but then cite them as the cause of the fire. This is logically incoherent. The fact that people do this speaks to the cognitive durability that misinformation can have [32,33].

The root of the matter appears to be that human memory does not simply write and overwrite like computer memory. Corrected misinformation does not vanish. Ecker et al. [15] trace this to the conflicting needs for memory stability and flexibility:

> Human memory is faced with the conundrum of maintaining stable memory representations (which is the whole point of having a memory in the first place) while also allowing for flexible modulation of memory representations to keep up-to-date with reality. Memory has evolved to achieve both of these aims, and hence it does not work like a blackboard: Outdated things are rarely actually wiped out and over-written; instead, they tend to linger in the background, and access to them is only gradually lost. [15] (p. 15)

There are some techniques for reducing the cognitive salience of misinformation; these are discussed in detail below. However, in many cases, it would be highly desirable to simply avoid the misinformation in the first place. Therefore, this section presents some strategies for preventing superintelligence misinformation.

The ideas for preventing superintelligence misinformation are inevitably more speculative than those for correcting it. There are two reasons for this. One is that the correction of misinformation has been the subject of a relatively extensive literature, while the prevention of misinformation has received fairly little scholarly attention. (Rare examples of studies on preventing misinformation are [34,35].) The other reason is that the correction of misinformation is largely cognitive and thus conducive to simple laboratory experiments, whereas the prevention of misinformation is largely sociological and thus requires a more complex and case-specific analysis. Nonetheless, given the importance of preventing superintelligence misinformation, it is important to consider potential strategies for doing so.

4.1. Educate Prominent Voices about Superintelligence

Perhaps the most straightforward approach to preventing superintelligence misinformation is to educate people who have prominent voices in discussions about superintelligence. The aim here is to give them a more accurate understanding of superintelligence so that they can pass that along to their respective audiences. Prominent voices about superintelligence can include select scholars, celebrities, or journalists, among others.

Educating the prominent may be easier said than done. For starters, they can be difficult to access, due to busy schedules and multitudes of other voices competing for their attention. Additionally, some of them they may already believe superintelligence misinformation, especially those who are already spreading it. Misinformation is difficult to correct in general, and may be even more difficult to correct for busy people who lack the mental attention to revise their thinking. (See Section 5.4 for further discussion of this point.) People already spreading misinformation may seem to be ideal candidates for educational efforts, in order to persuade them to change their tune, but it may actually be more productive to engage with people who have not yet made up their minds. Regardless, there is no universal formula for this sort of engagement, and the best opportunities may often be a matter of particular circumstance.

One model that may be of some value is the effort to improve the understanding of global warming among broadcast meteorologists. Broadcast meteorologists are for many people the primary messenger of environmental science. Furthermore, as a group, meteorologists (broadcast and non-broadcast) have traditionally been more skeptical about global warming than most of their peers in other Earth sciences [36,37]. In light of this, several efforts have been made to provide broadcast meteorologists with a better understanding of climate science, in hopes that they would pass this on to their audiences (e.g., [38,39]).

The case of broadcast meteorologists has important parallels to the many AI computer scientists who do not specialize in AGI or superintelligence. Both groups have expertise on a topic that is closely related to, but not quite the same as, the topic at hand. Broadcast meteorologists' expertise is weather, whereas global warming is about climate. (Weather concerns the day-to-day fluctuations in meteorological conditions, whereas climate concerns the long-term trends. An important distinction is that while weather can only be forecast a few days in advance, climate can be forecasted years or decades in advance.) Similarly, most AI computer scientists focus on AI that has "narrow" intelligence (intelligence in a limited range of domains), not AGI. Additionally, broadcast meteorologists and narrow AI computer scientists are often asked to voice their views on climate change and AGI, respectively.

4.2. Create Reputational Costs for Misinformers

When prominent voices cannot be persuaded to change their minds, they can at least be punished for getting it wrong. Legal punishment is possible in select cases (Section 4.5). However, reputational punishment is almost always possible and has potential to be quite effective, especially for public intellectuals whose brands depend on a good intellectual reputation.

In an analysis of US healthcare policy debates, Nyhan [40] concludes that correcting misinformation is extremely difficult and that increasing reputational costs may be more effective.

Nyhan [40] identifies misinformation that was critical to two healthcare debates: in the 1990s, the false claim that the policy proposed by President Bill Clinton would prevent people from keeping their current doctors, and in the 2000s, the false claim that the policy proposed by President Barack Obama would have established government "death panels" to deny life-sustaining coverage to the elderly. Nyhan [40] traces this misinformation to Betsy McCaughey, a scholar and politician generally allied with US conservative politics and opposed to these healthcare policy proposals:

> "Until the media stops giving so much attention to misinformers, elites on both sides will often succeed in creating misperceptions, especially among sympathetic partisans. And once such beliefs take hold, few good options exist to counter them—correcting misperceptions is simply too difficult. The most effective approach may therefore be for concerned scholars, citizens, and journalists to (a) create negative publicity for the elites who are promoting misinformation, increasing the costs of making false claims in the public sphere, and (b) pressure the media to stop providing coverage to serial dissemblers". [40] (p. 16)

Nyhan [40] further notes that while McCaughey's false claims were widely praised in the 1990s, including with a National Magazine Award, they were heavily criticized in the 2000s, damaging her reputation and likely reducing the spread of the misinformation.

There is some evidence indicating the possibility that reputational threats can succeed at reducing misinformation. Nyhan and Reifler [34] sent a randomized group of US state legislators a series of letters warning them about the reputational and electoral harms that the legislators could face if an independent fact checker (specifically, PolitiFact) finds them to make false statements. The study found that the legislators receiving the warnings were significantly less likely to make false statements. This finding is especially applicable to superintelligence misinformation spread by politicians, whose statements are more likely to be evaluated by fact checker like PolitiFact. Conceivably, similar fact checking systems could be developed for other types of public figures, or even for more low-profile professional discourse such as occurs among scientists and other technical experts. Similarly, Tsipursky and Morford [41] and Tsipursky et al. [35] describe a Pro-Truth Pledge aimed at committing people to refrain from spreading misinformation and to ask other people to retract misinformation, which can serve as a reputational punishment for misinformers, as well as a reputational benefit for those who present accurate information. Initial evaluations provide at least anecdotal support for the pledge having a positive effect on the information landscape.

For superintelligence misinformation, creating reputational costs has potential to be highly effective. A significant portion of influential voices in the debate have scholarly backgrounds and reputations that they likely wish to protect. For example, many of Steven Pinker's remarks about superintelligence are clearly misinformed, including the one discussed in Section 2 and several in his recent book *Enlightenment Now* [42]. (For detailed analysis of *Enlightenment Now*, see Torres [9].) Given Pinker's scholarly reputation, it may be productive to spread a message such as 'Steven Pinker is unenlightened about AI'.

At the same time, it is important to recognize the potential downsides of imposing reputational costs. Criticizing a person can damage one's relationship with them, reducing other sorts of opportunities. For example, criticizing people who may be building superintelligence could make them less receptive to other efforts to make their work safer. (Or, it could make them more receptive—this can be highly specific to individual personalities and contexts.) Additionally, it can impose reputational costs on the critic, such as a reputation of negativity or of seeking to restrict free speech. Caution is especially warranted for cases in which the misinformation comes from a professional contrarian, who may actually benefit from and relish in the criticism. For example, Marshall [43] (p. 72–73) warns climate scientists against debating professional climate deniers, since the latter tend to be more skilled at debate, especially televised debate, even though the arguments of the former are more sound. The same could apply for superintelligence, if it is to ever have a similar class of professional debaters. Thus, the imposition of reputation costs is a strategy to pursue selectively in certain instances of superintelligence misinformation.

4.3. Mobilize against Institutional Misinformation

The most likely institutional sources of superintelligence misinformation are the corporations involved in AI R&D, especially R&D for AGI and superintelligence. These companies have a vested interest in cultivating the impression that their technologies are safe and good for the world.

For these companies, reputational costs can also be significant. Corporate reputation can be important for consumer interest in the companies' products, citizen and government interest in imposing regulations on the companies, investor expectations of future profits, and employee interest in working for the companies. Therefore, one potential strategy is to incentivize companies so as to align their reputation with accurate information about superintelligence.

A helpful point of comparison is to corporate messaging about environmental issues, in particular the distinction between "greenwashing" and "brownwashing" [44]. Greenwashing is when a company portrays itself as protecting the environment when it is actually causing much environmental harm. For example, a fossil fuel company may publicize the greenhouse gas emissions reductions from solar panels it installs on its headquarters building while downplaying the fact that its core business model is a major driver of greenhouse gas emissions. In contrast, brownwashing is when a company declines to publicize its efforts towards environmental protection, perhaps because they have customers who oppose environmental protection or investors who worry it reduces profitability. In short, greenwashing is aimed at audiences that value environmental protection, while brownwashing is aimed at audiences that disvalue it.

Greenwashing is often criticized for giving companies a better environmental reputation than they deserve. In many cases that criticism may be fair. However, from an environmental communication standpoint, greenwashing does have the benefit of promoting a pro-environmental message. At a minimum, audiences of greenwashing are told that environmental protection is important. Audiences may also be given accurate information about environmental issues—for example, an advertisement that touts a fossil fuel company's greenhouse gas emissions reductions may also correctly explain that global warming is real and is caused by human action.

Similarly, there may be value in motivating AI companies to present accurate messages about superintelligence. This could be accomplished by cultivating demand for accurate messages among the companies' audiences. For example, if the public wants to hear accurate messages about superintelligence, then corporate advertising may be designed accordingly. The advertising might overstate the company's positive role, which would be analogous to greenwashing and could likewise be harmful for reducing accountability for bad corporate actors, but even then it would at least be spreading an accurate message about superintelligence.

Another strategy is for the employees of AI companies to mobilize against the companies supporting superintelligence misinformation, or against misinformation in general. At present, this may be a particularly promising strategy. There is a notable recent precedent for this in the successful employee action against Google's participation in Project Maven, a defense application of AI [45]. While not specifically focused on misinformation, this incident demonstrates the potential for employee action to change the practices of AI companies, including when those practices would otherwise be profitable for the company.

4.4. Focus Media Attention on Constructive Debates

Public media can inadvertently spread misinformation via the journalistic norm of balance. For the sake of objectivity, journalists often aim to cover "both sides" of an issue. While this can be constructive for some issues, it can also spread misinformation. For example, media coverage has often presented "both sides" of the "debate" over whether tobacco causes cancer or whether human activity causes global warming, even when one side is clearly correct and the other side has a clear conflict of interest [10,13].

One potential response for this is to attempt to focus media attention on legitimate open questions about a given issue, questions for which there are two meaningful sides to cover. For global warming, this could be a debate over the appropriate role of nuclear power or the merits of carbon taxes. For superintelligence, it could be a debate over the appropriate role of government regulations, or over the values that superintelligence (or AI in general) should be designed to promote. These sorts of debates satisfy the journalistic interest in covering two sides of an issue and provide a dramatic tension that can make for a better story, all while drawing attention to important open questions and affirming basic information about the topic.

4.5. Establish Legal Requirements

Finally, there may be some potential to legally require certain actors, especially corporations, to refrain from spreading misinformation. A notable precedent is the court decision of United States v. Philip Morris, in which nine tobacco companies and two tobacco trade organizations were found guilty of conspiring to deceive the public about the link between tobacco and cancer. Such legal decisions can have powerful effect.

However, legal requirements may be poorly suited to superintelligence misinformation. First, legal requirements can be slow to develop. The court case United States v. Philip Morris began in 1999, an initial ruling was reached in 2006, and that ruling was upheld in 2009. Furthermore, United States v. Philip Morris came only after several decades of tobacco industry misinformation. Given the evolving nature of AI technology, it could be difficult to pin down which information is correct over such long time periods. Second, superintelligence is a future technology for which much of the correct information cannot be established with the same degree of rigor. Furthermore, if and when superintelligence is built, it could be so transformative as to render current legal systems irrelevant. (For more general discussion of the applicability of legal mechanisms to superintelligence, see [46–48].) For these reasons, legal requirements are less likely to play a significant role in preventing superintelligence misinformation.

5. Correcting Superintelligence Misinformation

Correcting misinformation is sufficiently difficult that it will often be better to prevent it from spreading in the first place. However, when superintelligence misinformation cannot be prevented, there are strategies available for correcting it in the minds of those who are exposed to it. Correcting misinformation is the subject of a fairly extensive literature in psychology, political science, and related fields [2,15,33,49]. For readers unfamiliar with this literature, Cook et al. [2] provide an introductory overview accessible to an interdisciplinary readership, while Ecker et al. [15] provide a more detailed and technical survey. This section applies this literature to the correction of superintelligence misinformation.

5.1. Build Expert Consensus and the Perception Thereof

At present, there exists substantial expert disagreement about a wide range of aspects of superintelligence, from basic matters such as whether superintelligence is possible [50–52] and when it might occur if it does [53–55] to subtler matters such as the treacherous turn vs. the sordid stumble. The situation stands in contrast to the extensive expert consensus on other issues such as global warming [56]. (Experts lack consensus on some important details about global warming, such as how severe the damage is likely to be, but they have a high degree of consensus on the basic contours of the issue.).

The case of global warming shows that expert consensus on its own does not counteract misinformation. On the contrary, misinformation about global warming continues to thrive despite the existence of consensus. However, there is reason to believe that the consensus helps. For starters, much of the misinformation is specifically oriented towards creating the false perception that there is no consensus [10]. The scientific consensus is a target of misinformation because it is believed to be an important factor in people's overall beliefs. Indeed, several studies have documented a strong correlation among the lay public between rejection of the science of global warming and belief that there is no consensus [57,58]. Further studies find that presenting messages describing the consensus increases belief in climate science and support for policy to reduce greenhouse gas emissions [14,59]. Notably, this effect is observed for people across the political spectrum, including those who would have political motivation to doubt the science. (Such motivations are discussed further in Section 5.2.) All of this indicates an important role for expert consensus in broader beliefs about global warming.

For superintelligence, at present there is no need to spread misinformation about the existence of consensus because there is rather little consensus. Therefore, a first step is to work towards consensus. (This of course should be consensus grounded on the best possible analysis, not consensus for the sake of consensus.) This may be difficult for superintelligence because of the inherent challenge of understanding future technologies and the complexity of advanced AI. Global warming has its own complexities, but the core science is relatively simple: increased atmospheric greenhouse gas concentrations trap sunlight and raise temperatures. However, at least some aspects of superintelligence should be easy enough to get consensus on, starting with the fact that there are a number of R&D groups attempting to build AGI. Other aspects may be more difficult to build consensus on, but this consensus is at least something that can be pursued via normal channels of expert communication: research articles, conference symposia, private correspondence, and so on.

Given the existence of consensus, it is also important to raise awareness about it. The consensus cannot counteract misinformation if nobody knows about it. The global warming literature provides good models for documenting expert consensus [56], and such findings of consensus can be likewise be publicized.

5.2. Address Pre-Existing Motivations for Believing Misinformation

The human mind tends to not process new information in isolation, but instead processes it in relation to wider beliefs and understandings of the world. This can be very valuable, enabling us to understand the context behind new information and relate it to existing knowledge. For example, people would typically react with surprise and confusion upon seeing an object rise up to the ceiling instead of fall down to the floor. This new information is related to a wider understanding of the fact that objects fall downwards. People may even struggle to believe their own eyes unless there is a compelling explanation. (For example, perhaps the object and the ceiling are both magnetized.). Additionally, if people did not see it with their own eyes, but instead heard it reported by someone else, they may be even less likely to believe it. In other words, they are motivated to believe that the story is false, even if it is true. This phenomenon is known as *motivated reasoning*.

While generally useful, motivated reasoning can be counterproductive in the context of misinformation, prompting people to selectively believe misinformation over correct information. This occurs in particular when the misinformation accords better with preexisting beliefs than the correct information. In the above example, misinformation could be that the object fell down to the floor instead of rising to the ceiling.

Motivated reasoning is a major factor in the belief of misinformation about politically contentious issues such as climate change. The climate science consensus is rejected mainly by people who believe that government regulation of industry is generally a bad thing [14,59]. In principle, belief that humans are warming the planet should have nothing to do with belief that government regulations are harmful. It is logically coherent to believe in global warming yet argue that carbon emissions should not be regulated. However, in practice, the science of global warming often threatens people's wider beliefs about regulations, and so they find themselves motivated to reject the science.

Motivated reasoning can also be a powerful factor for beliefs about superintelligence. A basic worldview is that humans are in control. Per this worldview, human technology is a tool; the idea that it could rise up against humanity is a trope for science fiction, not something to be taken seriously. The prospect of superintelligence threatens this worldview, predisposing people to not take superintelligence seriously. In this context, it may not help that media portrayals of the scholarly debate about superintelligence commonly include reference to science fiction, such as by using pictures of the Terminator. As one expert who is concerned about superintelligence states, "I think that at this point all of us on all sides of this issue are annoyed with the journalists who insist on putting a picture of the Terminator on every single article they publish of this topic" [60].

Motivated reasoning has been found to be linked to people's sense of self-worth. As one study puts it, "the need for self-integrity—to see oneself as good, virtuous, and efficacious—is a basic human motivation" [61] (p. 415). When correct information threatens people's self-worth, they are more motivated to instead believe misinformation, so as to preserve their self-worth. Furthermore, motivated reasoning can be reduced by having people consciously reaffirm their own self-worth, such as by recalling to themselves ways in which they successfully live up to their personal values [61]. Essentially, with their sense of self-worth firmed up, they become more receptive to information that would otherwise threaten their self-worth.

As a technology that could outperform humans, superintelligence could pose an especially pronounced threat to people's sense of self-worth. It may be difficult for people to feel good and efficacious if they would soon be superseded by computers. For at least some people, this could be a significant reason to reject information about the prospect of superintelligence, even if that information is true. At the same time, it may still be valuable for messages about superintelligence to be paired with messages of affirmation.

Another important set of motivations comes from the people active in superintelligence debates. Many people in the broader computer science field of AI have been skeptical of claims about superintelligence. These people may be motivated by a desire to protect the reputation and funding of the field of AI, and in turn protect their self-worth as AI researchers. AI has a long history of boom-bust cycles in which hype about superintelligence and related advanced AI falls flat and contributes to an "AI winter". Peter Bentley, an AI computer scientist who has spoken out against contemporary claims about superintelligence, is explicit about this:

> "Large claims lead to big publicity, which leads to big investment, and new regulations. And then the inevitable reality hits home. AI does not live up to the hype. The investment dries up. The regulation stifles innovation. And AI becomes a dirty phrase that no-one dares speak. Another AI Winter destroys progress" [62] (p. 10). "Do not be fearful of AI—marvel at the persistence and skill of those human specialists who are dedicating their lives to help create it. And appreciate that AI is helping to improve our lives every day" (p. 11).

While someone's internal motivations can only be inferred from such text, the text is at least suggestive of motivations to protect self-worth and livelihood as an AI researcher, as well as a worldview in which AI is a positive force for society.

To take another example, Torres [9] proposes that Pinker's dismissal of AGI and superintelligence is motivated by Pinker's interest in promoting a narrative in which science and technology bring progress—a narrative that could be threatened by the potential catastrophic risk from superintelligence.

Conversely, some people involved in superintelligence debates may be motivated to believe in the prospect of superintelligence. For example, researcher Jürgen Schmidhuber writes on his website that "since age 15 or so, the main goal of professor Jürgen Schmidhuber has been to build a self-improving Artificial Intelligence (AI) smarter than himself, then retire." [63] Superintelligence is also sometimes considered the "grand dream" of AI [64]. Other common motivations include a deep interest in transformative future outcomes [65] and a deep concern about extreme catastrophic risks [4,66,67]. People with these worldviews may be predisposed to believe certain types of claims about superintelligence. If it turns out that superintelligence will not be built, or would not have transformative or catastrophic effects, then this can undercut people's deeply held beliefs in the importance of superintelligence, transformative futures, and/or catastrophic risks.

For each of these motivations for interest in superintelligence, there can be information that is rejected because it cuts against the motivations and misinformation that is accepted because it supports the motivations. Therefore, in order to advance superintelligence debates, it can be valuable to affirm people's motivations when presenting conflicting information. For example, one could affirm that

AI computer scientists are making impressive and important contributions to the world, and then explain reasons why superintelligence may nonetheless be a possibility worth considering. One could affirm that science and technology are bringing a great deal of progress, and then explain reasons why some technologies could nonetheless be dangerous. One could affirm that superintelligence is indeed a worthy dream, or that transformative futures are indeed important to pay attention to, and then explain reasons why superintelligence might not be built. Finally, one could affirm that extreme catastrophic risks are indeed an important priority for human society, and then explain reasons why superintelligence may not be such a large risk after all. These affirming messaging strategies could predispose participants in superintelligence debates to consider a wider range of possibilities and make more progress on the issue, including progress towards expert consensus.

Another strategy is to align motivations with accurate beliefs about superintelligence. For example, some AI computer scientists may worry that belief in the possibility of superintelligence could damage reputation and funding. However, if belief in the possibility of superintelligence would bring reputational and funding benefits, then the same people may be more comfortable expressing such belief. Reputational benefits could be created, for example, via slots in high-profile conferences and journals, or by association with a critical mass of reputable computer scientists who also believe in the possibility of superintelligence. Funding could likewise be made available. Noting that funding and space in conferences and journals are often scarce resources, it could be advantageous to target these resources at least in part toward shifting motivations of important actors in superintelligence debates. This example of course assumes that it is correct to believe in the possibility of superintelligence. The same general strategy of aligning motivations may likewise be feasible for other beliefs about superintelligence.

The above examples—concerning the reputations and funding of AI computer scientists, the possibility of building superintelligence, and the importance of transformative futures and catastrophic risks—all involve experts or other communities that are relatively attentive to the prospect of superintelligence. Other motivations could be significant for the lay public, policy makers, and other important actors. Research on the public understanding of science finds that cultural factors, such as political ideology, can factor significantly in the interpretation of scientific information [68,69]. Kahan et al. [69] (p. 79) propose to "shield" scientific evidence and related information "from antagonistic cultural information". For superintelligence, this could mean attempting to frame superintelligence (or, more generally, AI) as a nonpartisan social issue. At least in the US, if an issue becomes politically partisan, legislation typically becomes substantially more difficult to pass. Likewise, discussions of AI and superintelligence should, where reasonably feasible, attempt to avoid close association with polarizing ideologies and cultural divisions.

The fact that early US legislation on AI has been bipartisan is encouraging. For example, H.R.4625, FUTURE of Artificial Intelligence Act of 2017, sponsored by John Delaney (Democrat) and co-sponsored by Pete Olson (Republican), and H.R.5356, National Security Commission Artificial Intelligence Act of 2018, sponsored by Elise Stefanik (Republican) and co-sponsored by James Langevin (Democrat). This is a trend that should be praised and encouraged to continue.

5.3. Inoculate with Advance Warnings

The misinformation literature has developed the concept of *inoculation*, in which people are preemptively educated about a piece of misinformation so that they will not believe it if and when they later hear it. For example, someone might be told that there is a false rumor that vaccines cause autism, such that when they later hear the rumor, they know to recognize it as false. The aim is to get people to correctly understand the truth about a piece of misinformation from the beginning, so that their minds never falsely encode it. Inoculation has been found to work better than simply telling people the correct information [70].

Inoculation messages can include why a piece of misinformation is incorrect as well as why it is being spread [71]. For example, misinformation casting doubt on the idea that global temperatures are rising could be inoculated with an explanation of how scientists have established that global temperatures are rising. The inoculation could also explain that industries are intentionally casting doubt about global temperature increases in order to avoid regulations and increase profits. Likewise, for superintelligence, misinformation claiming that there are no projects seeking to build AGI could be inoculated by explanations of the existence of AGI R&D projects, and perhaps also explanations of the motivations of people who claim that there are no such projects. For example, Torres [9] proposes that Pinker's dismissal of AGI and superintelligence is motivated by Pinker's interest in promoting a narrative in which science and technology bring progress—a narrative that could be threatened by the potential catastrophic risk from superintelligence.

5.4. Explain Misinformation and Corrections

When people are exposed to misinformation, it can be difficult to correct, as first explained in Section 4. This phenomenon has been studied in great depth, with the terms "continued influence" and "belief perseverance" used for cases in which debunked information continues to influence people's thinking [72,73]. There is also an "illusion of truth", in which information explained to be false is later misremembered as true—essentially, the mind remembers the information but forgets its falsity [74]. The difficulty of correcting misinformation is why this paper has emphasized strategies to prevent of misinformation from spreading in the first place.

Adding to the challenge is the fact that attempts to debunk misinformation can inadvertently reinforce it. This phenomenon is known as the "backfire effect" [74]. Essentially, when someone hears "X is false", it can strengthen their mental representation of X, thereby reinforcing the misinformation. This effect has been found to be especially pronounced among the elderly [74]. One explanation is that correcting the misinformation (i.e., successfully processing "X is false") requires the use of strategic memory, but strategic memory requires dedicated mental effort and is less efficient among the elderly [15]. Unless enough strategic memory is allocated to processing "X is false", the statement can end up reinforcing belief in X.

These findings about the backfire effect have important consequences for superintelligence misinformation. Fortunately, many important audiences for superintelligence misinformation are likely to have strong strategic memories. Among the prominent actors in superintelligence debates, relatively few are elderly, and many of them have intellectual pedigrees that may endow them with strong strategic memories. On the other hand, many of the prominent actors are busy people with limited mental energy available for processing corrections about superintelligence information. As a practical matter, people attempting to debunk superintelligence misinformation should generally avoid "X is false" messages, especially when their audience may be paying limited attention.

One technique that has been particularly successful at correcting misinformation is the use of refutational text, which provides detailed explanations of why the misinformation is incorrect, what the correct information is, and why it is correct. Refutational text has been used mainly as a classroom tool for helping students overcome false preexisting beliefs about course topics [75,76]. Refutational text has even been used to turn misinformation into a valuable teaching tool [77]. A meta-analysis found refutational text to be the most effective technique for correcting misinformation in the context of science education—that is, for enabling students to overcome preexisting misconceptions about science topics [78].

A drawback of refutational text is that it can require more effort and attention than simpler techniques. Refutational text may be a valuable option in classrooms or other settings in which one has an audience's extended attention. Such settings include many venues of scholarly communication, which can be important for superintelligence debates. However, refutational texts may be less viable in other settings, such as social media and television news program interviews, in which one can often only get in a short sound bite. Therefore, refutational text may be relatively well-suited for interactions with experts and other highly engaged participants in superintelligence debates, and relatively poorly suited for much of the lay public and others who may only hear occasional passing comments about superintelligence. That said, it may still be worth producing and disseminating extended refutations for lay public audiences, such as in long-format videos and articles for television, magazines, and online. These may tend to only reach the most motivated segments of the lay public, but they can nonetheless be worthwhile.

6. Conclusions

Superintelligence is a high-stakes potential future technology as well as a highly contested socio-technological issue. It is also fertile terrain for misinformation. Making progress on the issue requires identifying and rejecting misinformation and accepting accurate information. Some progress will require technical research to clarify the nature of superintelligence. However, a lot of progress will likely also require the sorts of sociological and psychological strategies outlined in this paper. The most progress may come from interdisciplinary projects connecting computer science, social science, and other relevant fields. Computer science is a highly technical field, but as with all fields, it is ultimately composed of human beings. By appreciating the nuances of the human dimensions of the field, it may be possible to make better progress towards understanding superintelligence and acting responsibly about it.

As the first dedicated study of strategies for countering superintelligence misinformation, this paper has taken a broad view, surveying a range of options. Despite this breadth, there may still be additional options worth further attention. Indeed, this paper has only mined a portion of the insights contained within the existing literature on misinformation. There may also be compelling options that go beyond the literature. Likewise, because of this paper's breadth, it has given relatively shallow treatment to each of the options. More detailed attention to the various option would be another worthy focus of future research.

An especially valuable focus would be the proposed strategies for preventing superintelligence misinformation. Because misinformation can be so difficult to correct, preventing it may be the more effective strategy. There is also less prior research on the prevention of misinformation. For these reasons, there is likely to be an abundance of important research opportunities on the prevention of misinformation, certainly for superintelligence misinformation and perhaps also for misinformation in general.

For the prevention of superintelligence misinformation, a strategy that may be particularly important to study further is dissuading AI corporations from using their substantial resources to spread superintelligence misinformation. The long history of corporations engaging in such tactics, with a major impact on the surrounding debates, suggests that this could be a highly important factor for superintelligence [12]. It may be especially valuable to study this at an early stage, before such tactics are adopted.

For the correction of superintelligence misinformation, a particularly promising direction is on the motivations and worldviews of prominent actors and audiences in superintelligence debates. Essentially, what are people's motivations with respect to superintelligence? Are AI experts indeed motivated to protect their field? Are superintelligence developers motivated by the "grand dream"? Are others who believe in the prospect of superintelligence motivated by beliefs about transformative futures or catastrophic risks? Can attention to these sorts of motivations help them overcome their divergent worldviews and make progress towards consensus on the topic? Finally, are people in general motivated to retain their sense of self-worth in the face of a technology that could render them inferior?

Most important, however, is not the research on superintelligence misinformation, but the efforts to prevent and correct it. It can often be stressful and thankless work, especially amidst the heated debates, but it is essential to ensuring positive outcomes. This paper is one effort towards helping this work succeed. Given the exceptionally high potential stakes, it is vital that decisions about superintelligence be well-informed.

Funding: This research received no external funding.

Acknowledgments: Olle Häggström, Tony Barrett, Brendan Nyhan, Maurizio Tinnirello, Stephan Lewandowsky, Michael Laakasuo, Phil Torres, and three anonymous reviewers provided helpful feedback on earlier versions of this paper. All remaining errors are the author's alone. The views expressed in this paper are the author's and not necessarily the views of the Global Catastrophic Risk Institute.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Definition of Misinformation in English by Oxford Dictionaries. Available online: https://en.oxforddictionaries.com/definition/misinformation (accessed on 9 September 2018).
2. Cook, J.; Ecker, U.; Lewandowsky, S. Misinformation and how to correct it. In *Emerging Trends in the Social and Behavioral Sciences*; John Wiley & Sons: Hoboken, NJ, USA, 2015; pp. 1–17.
3. Kelly, K. The Myth of a Superhuman AI. Available online: https://www.wired.com/2017/04/the-myth-of-a-superhuman-ai (accessed on 24 September 2018).
4. Häggström, O. *Here Be Dragons: Science, Technology and the Future of Humanity*; Oxford University Press: Oxford, UK, 2016.
5. Häggström, O. Michael Shermer Fails in His Attempt to Argue That AI Is Not an Existential Threat. *Häggström Hävdar*. 19 September 2017. Available online: http://haggstrom.blogspot.com/2017/09/michael-shermer-fails-in-his-attempt-to.html (accessed on 24 September 2018).
6. Häggström, O. The AI meeting in Brussels Last Week. *Häggström Hävdar*. 23 October 2017. Available online: http://haggstrom.blogspot.com/2017/10/the-ai-meeting-in-brussels-last-week.html (accessed on 9 September 2018).
7. Muehlhauser, L. *Three Misconceptions in Edge.org's Conversation on "The Myth of AI"*; Machine Intelligence Research Institute: Berkeley, CA, USA, 18 November 2014; Available online: https://intelligence.org/2014/11/18/misconceptions-edge-orgs-conversation-myth-ai (accessed on 24 September 2018).
8. Torres, P. Why Superintelligence Is a Threat That Should Be Taken Seriously. *Bulletin of the Atomic Scientists*. 24 October 2017. Available online: https://thebulletin.org/why-superintelligence-threat-should-be-taken-seriously11219 (accessed on 24 September 2018).
9. Torres, P. A Detailed Critique of One Section of Steven Pinker's Chapter "Existential Threats" in Enlightenment Now. Project for Future Human Flourishing Technical Report 2, Version 1.2. 2018. Available online: https://docs.wixstatic.com/ugd/d9aaad_8b76c6c86f314d0288161ae8a47a9821.pdf (accessed on 9 September 2018).
10. Oreskes, N.; Conway, E.M. *Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming*; Bloomsbury: New York, NY, USA, 2010.
11. Grandjean, P. *Only One Chance: How Environmental Pollution Impairs Brain Development—And How to Protect the Brains of the Next Generation*; Oxford University Press: Oxford, UK, 2013.
12. Baum, S.D. Superintelligence skepticism as a political tool. *Information* **2018**, *9*, 209. [CrossRef]
13. Boykoff, M.T.; Boykoff, J.M. Balance as bias: Global warming and the US prestige press. *Glob. Environ. Chang.* **2004**, *14*, 125–136. [CrossRef]
14. Lewandowsky, S.; Gignac, G.E.; Vaughan, S. The pivotal role of perceived scientific consensus in acceptance of science. *Nat. Clim. Chang.* **2013**, *3*, 399–404. [CrossRef]
15. Ecker, U.K.H.; Swire, B.; Lewandowsky, S. Correcting misinformation—A challenge for education and cognitive science. In *Processing Inaccurate Information: Theoretical and Applied Perspectives from Cognitive Science and the Educational Sciences*; Rapp, D.N., Braasch, J.L.G., Eds.; MIT Press: Cambridge, MA, USA, 2014; pp. 13–38.
16. Baum, S.D. On the promotion of safe and socially beneficial artificial intelligence. *AI Soc.* **2017**, *32*, 543–551. [CrossRef]
17. Lewandowsky, S.; Ecker, U.K.H.; Seifert, C.M.; Schwarz, N.; Cook, J. Misinformation and its correction: Continued influence and successful debiasing. *Psychol. Sci. Public Interest* **2012**, *13*, 106–131. [CrossRef] [PubMed]

18. Pinker, S. We're Told to Fear Robots. But Why Do We Think They'll Turn on Us?". *Popular Science*. 13 February 2018. Available online: https://www.popsci.com/robot-uprising-enlightenment-now (accessed on 9 September 2018).
19. Goertzel, B. Artificial general intelligence: Concept, state of the art, and future prospects. *J. Artif. Gen. Intell.* **2014**, *5*, 1–48. [CrossRef]
20. Baum, S.D. A Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy. Global Catastrophic Risk Institute Working Paper 17-1. 2017. Available online: https://ssrn.com/abstract=3070741 (accessed on 9 September 2018).
21. Cognitive Artificial Intelligence: The MicroPsi Project. Available online: http://cognitive-ai.com (accessed on 9 September 2018).
22. Searle, J.R. What Your Computer Can't Know. *New York Review of Books*, 9 October 2014.
23. Logan, R.K. Can computers become conscious, an essential condition for the Singularity? *Information* **2017**, *8*, 161. [CrossRef]
24. Chalmers, D.J. The singularity: A philosophical analysis. *J. Conscious. Stud.* **2010**, *17*, 7–65.
25. Bostrom, N. *Superintelligence: Paths, Dangers, Strategies*; Oxford University Press: Oxford, UK, 2014.
26. Omohundro, S.M. The basic AI drives. In *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*; Wang, P., Goertzel, B., Franklin, S., Eds.; IOS: Amsterdam, The Netherlands, 2008; pp. 483–492.
27. Goertzel, B. Infusing advanced AGIs with human-like value systems: Two theses. *J. Evol. Technol.* **2016**, *26*, 50–72.
28. Baum, S.D.; Barrett, A.M.; Yampolskiy, R.V. Modeling and interpreting expert disagreement about artificial superintelligence. *Informatica* **2017**, *41*, 419–428.
29. Danaher, J. Why AI doomsayers are like sceptical theists and why it matters. *Minds Mach.* **2015**, *25*, 231–246. [CrossRef]
30. Hughes, J.J. Global technology regulation and potentially apocalyptic technological threats. In *Nanoethics: The Ethical and Social Implications of Nanotechnology*; Allhof, F., Ed.; Wiley: Hoboken, NJ, USA, 2007; pp. 201–214.
31. Yampolskiy, R.; Fox, J. Safety engineering for artificial general intelligence. *Topoi* **2013**, *32*, 217–226. [CrossRef]
32. Wilkes, A.L.; Leatherbarrow, M. Editing episodic memory following the identification of error. *Q. J. Exp. Psychol.* **1988**, *40A*, 361–387. [CrossRef]
33. Johnson, H.M.; Seifert, C.M. Sources of the continued influence effect: When misinformation in memory affects later inferences. *J. Exp. Psychol. Learn. Mem. Cognit.* **1994**, *20*, 1420–1436. [CrossRef]
34. Nyhan, B.; Reifler, J. The effect of fact-checking on elites: A field experiment on U.S. state legislators. *Am. J. Political Sci.* **2015**, *59*, 628–640. [CrossRef]
35. Tsipursky, G.; Votta, F.; Roose, K.M. Fighting fake news and post-truth politics with behavioral science: The pro-truth pledge. *Behav. Soc. Issues* **2018**, *27*, 47–70. [CrossRef]
36. Doran, P.T.; Zimmerman, M.K. Examining the scientific consensus on climate change. *Eos* **2009**, *90*, 22–23. [CrossRef]
37. Stenhouse, N.; Maibach, E.; Cobb, S.; Ban, R.; Bleistein, A.; Croft, P.; Bierly, E.; Seitter, K.; Rasmussen, G.; Leiserowitz, A. Meteorologists' views about global warming: A survey of American Meteorological Society professional members. *Bull. Am. Meteorol. Soc.* **2014**, *95*, 1029–1040. [CrossRef]
38. De La Harpe, J. TV Meteorologists, Weathercasters Briefed by Climate Experts at AMS Short Course. *Yale Climate Connnections*. 9 July 2009. Available online: https://www.yaleclimateconnections.org/2009/07/tv-meteorologists-weathercasters-briefedby-climate-experts-at-ams-short-course (accessed on 9 September 2018).
39. Ward, B. 15 Midwest TV Meteorologists, Weathercasters Weigh Climate Science at Chicago's Field Museum Climate Science for Meteorologists. *Yale Climate Connnections*. 5 May 2009. Available online: https://www.yaleclimateconnections.org/2009/05/meteorologists-weathercasters-weigh-climate-science-chicago (accessed on 9 September 2018).
40. Nyhan, B. Why the 'death panel' myth wouldn't die: Misinformation in the health care reform debate. *The Forum* **2010**, *8*. [CrossRef]
41. Tsipursky, G.; Morford, Z. Addressing behaviors that lead to sharing fake news. *Behav. Soc. Issues* **2018**, *27*, AA6–AA10.

42. Pinker, S. *Enlightenment Now: The Case for Reason, Science, Humanism, and Progress*; Penguin: New York, NY, USA, 2018.
43. Marshall, G. *Don't Even Think about It: Why Our Brains Are Wired to Ignore Climate Change*; Bloomsbury: New York, NY, USA, 2014.
44. Kim, E.-H.; Lyon, T.P. Greenwash vs. brownwash: Exaggeration and undue modesty in corporate sustainability disclosure. *Organ. Sci.* **2014**, *26*, 705–723. [CrossRef]
45. BBC. Google 'to end' Pentagon Artificial Intelligence Project. *BBC*. 2 June 2018. Available online: https://www.bbc.com/news/business-44341490 (accessed on 9 September 2018).
46. McGinnis, J.O. Accelerating AI. *Northwest. Univ. Law Rev.* **2010**, *104*, 366–381.
47. Wilson, G. Minimizing global catastrophic and existential risks from emerging technologies through international law. *Va. Environ. Law J.* **2013**, *31*, 307–364.
48. White, T.N.; Baum, S.D. Liability law for present and future robotics technology. In *Robot Ethics 2.0*; Lin, P., Abney, K., Jenkins, R., Eds.; Oxford University Press: Oxford, UK, 2017; pp. 66–79.
49. Ecker, U.K.H.; Lewandowsky, S.; Tang, D.T.W. Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Mem. Cognit.* **2010**, *38*, 1087–1100. [CrossRef] [PubMed]
50. Bringsjord, S. Belief in the singularity is logically brittle. *J. Conscious. Stud.* **2012**, *19*, 14–20.
51. McDermott, D. Response to the singularity by David Chalmers. *J. Conscious. Stud.* **2012**, *19*, 167–172.
52. Chalmers, D. The Singularity: A reply. *J. Conscious. Stud.* **2012**, *19*, 141–167.
53. Baum, S.D.; Goertzel, B.; Goertzel, T.G. How long until human-level AI? Results from an expert assessment. *Technol. Forecast. Soc. Chang.* **2011**, *78*, 185–195. [CrossRef]
54. Müller, V.C.; Bostrom, N. Future progress in artificial intelligence: A survey of expert opinion. In *Fundamental Issues of Artificial Intelligence*; Müller, V.C., Ed.; Springer: Cham, Switzerland, 2016; pp. 555–572.
55. Grace, K.; Salvatier, J.; Dafoe, A.; Zhang, B.; Evans, O. When will AI exceed human performance? Evidence from AI experts. *J. Artif. Intell. Res.* **2018**, *62*, 729–754. [CrossRef]
56. Oreskes, N. The scientific consensus on climate change. *Science* **2004**, *306*, 1686. [CrossRef] [PubMed]
57. Ding, D.; Maibach, E.W.; Zhao, X.; Roser-Renouf, C.; Leiserowitz, A. Support for climate policy and societal action are linked to perceptions about scientific agreement. *Nat. Clim. Chang.* **2011**, *1*, 462–466. [CrossRef]
58. McCright, A.M.; Dunlap, R.E.; Xiao, C. Perceived scientific agreement and support for government action on climate change in the USA. *Clim. Chang.* **2013**, *119*, 511–518. [CrossRef]
59. Van der Linden, S.L.; Leiserowitz, A.A.; Feinberg, G.D.; Maibach, E.W. The scientific consensus on climate change as a gateway belief: Experimental evidence. *PLoS ONE* **2015**, *10*, e0118489. [CrossRef] [PubMed]
60. Bensinger, R. Sam Harris and Eliezer Yudkowsky on 'AI: Racing toward the Brink'. *Machine Intelligence Research Institute*. 28 February 2018. Available online: https://intelligence.org/2018/02/28/sam-harris-and-eliezer-yudkowsky (accessed on 9 September 2018).
61. Cohen, G.L.; Sherman, D.K.; Bastardi, A.; Hsu, L.; McGoey, M.; Ross, L. Bridging the partisan divide: Self-affirmation reduces ideological closed-mindedness and inflexibility in negotiation. *J. Pers. Soc. Psychol.* **2007**, *93*, 415–430. [CrossRef] [PubMed]
62. Bentley, P.J. The three laws of artificial intelligence: Dispelling common myths. In *Should We Fear Artificial Intelligence? In-Depth Analysis*; Boucher, P., Ed.; European Parliamentary Research Service, Strategic Foresight Unit: Brussels, Belgium, 2018; pp. 6–12.
63. Jürgen Schmidhuber's Home Page. Available online: http://people.idsia.ch/~juergen (accessed on 9 September 2018).
64. Legg, S. Machine Super Intelligence. Doctoral's Thesis, University of Lugano, Lugano, Switzerland, 2008.
65. More, M.; Vita-More, N. (Eds.) *The Transhumanist Reader: Classical and Contemporary Essays on the Science, Technology, and Philosophy of the Human Future*; Wiley: New York, NY, USA, 2010.
66. Bostrom, N. Existential risk prevention as global priority. *Glob. Policy* **2013**, *4*, 15–31. [CrossRef]
67. Torres, P. *Morality, Foresight & Human Flourishing an Introduction to Existential Risks*; Pitchstone Publishing: Durham, NC, USA, 2017.
68. Kahan, D.M.; Jenkins-Smith, H.; Braman, D. Cultural cognition of scientific consensus. *J. Risk Res.* **2011**, *14*, 147–174. [CrossRef]
69. Kahan, D.M.; Peters, E.; Dawson, E.C.; Slovic, P. Motivated numeracy and enlightened self-government. *Behav. Public Policy* **2017**, *1*, 54–86. [CrossRef]

70. Banas, J.A.; Rains, S.A. A meta-analysis of research on inoculation theory. *Commun. Monogr.* **2010**, *77*, 281–311. [CrossRef]
71. Cook, J.; Lewandowsky, S.; Ecker, U.K.H. Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. *PLoS ONE* **2017**, *12*, e0175799. [CrossRef] [PubMed]
72. Cobb, M.D.; Nyhan, B.; Reifler, J. Beliefs don't always persevere: How political figures are punished when positive information about them is discredited. *Political Psychol.* **2013**, *34*, 307–326. [CrossRef]
73. Nyhan, B.; Reifler, J. Displacing misinformation about events: An experimental test of causal corrections. *J. Exp. Political Sci.* **2015**, *2*, 81–93. [CrossRef]
74. Skurnik, I.; Yoon, C.; Park, D.C.; Schwarz, N. How warnings about false claims become recommendations. *J. Consum. Res.* **2005**, *31*, 713–724. [CrossRef]
75. Kowalski, P.; Taylor, A.K. The effect of refuting misconceptions in the introductory psychology class. *Teach. Psychol.* **2009**, *36*, 153–159. [CrossRef]
76. Kuhn, D.; Crowell, A. Dialogic argumentation as a vehicle for developing young adolescents' thinking. *Psychol. Sci.* **2011**, *22*, 545–552. [CrossRef] [PubMed]
77. Bedford, D. Agnotology as a teaching tool: Learning climate science by studying misinformation. *J. Geogr.* **2010**, *109*, 159–165. [CrossRef]
78. Guzzetti, B.J.; Snyder, T.E.; Glass, G.V.; Gamas, W.S. Promoting conceptual change in science: A comparative meta-analysis of instructional interventions from reading education and science education. *Read. Res. Q.* **1993**, *28*, 117–159. [CrossRef]

information

MDPI

Article

Superintelligence Skepticism as a Political Tool

Seth D. Baum

Global Catastrophic Risk Institute, P.O. Box 40364, Washington, DC 20016, USA; seth@gcrinstitute.org

Received: 27 June 2018; Accepted: 17 August 2018; Published: 22 August 2018

Abstract: This paper explores the potential for skepticism about artificial superintelligence to be used as a tool for political ends. Superintelligence is AI that is much smarter than humans. Superintelligence does not currently exist, but it has been proposed that it could someday be built, with massive and potentially catastrophic consequences. There is substantial skepticism about superintelligence, including whether it will be built, whether it would be catastrophic, and whether it is worth current attention. To date, superintelligence skepticism appears to be mostly honest intellectual debate, though some of it may be politicized. This paper finds substantial potential for superintelligence skepticism to be (further) politicized, due mainly to the potential for major corporations to have a strong profit motive to downplay concerns about superintelligence and avoid government regulation. Furthermore, politicized superintelligence skepticism is likely to be quite successful, due to several factors including the inherent uncertainty of the topic and the abundance of skeptics. The paper's analysis is based on characteristics of superintelligence and the broader AI sector, as well as the history and ongoing practice of politicized skepticism on other science and technology issues, including tobacco, global warming, and industrial chemicals. The paper contributes to literatures on politicized skepticism and superintelligence governance.

Keywords: artificial intelligence; superintelligence; skepticism

1. Introduction

The purpose of this paper is to explore the potential for skepticism about artificial superintelligence to be used for political ends. Artificial superintelligence (for brevity, henceforth just *superintelligence*) refers to AI that is much smarter than humans. Current AI is not superintelligent, but the prospect of superintelligence is a topic of much discussion in scholarly and public spheres. Some believe that superintelligence could someday be built, and that, if it is built, it would have massive and potentially catastrophic consequences. Others are skeptical of these beliefs. While much of the existing skepticism appears to be honest intellectual debate, there is potential for it to be politicized for other purposes.

In simple terms (to be refined below), *politicized skepticism* can be defined as public articulation of skepticism that is intended to achieve some outcome other than an improved understanding of the topic at hand. Politicized skepticism can be contrasted with *intellectual skepticism*, which seeks an improved understanding. Intellectual skepticism is essential to scholarly inquiry; politicized skepticism is not. The distinction between the two is not always clear; statements of skepticism may have both intellectual and political motivations. The two concepts can nonetheless be useful for understanding debates over issues such as superintelligence.

There is substantial precedent for politicized skepticism. Of particular relevance for superintelligence is politicized skepticism about technologies and products that are risky but profitable, henceforth *risk–profit politicized skepticism*. This practice dates to 1950s debates over the link between tobacco and cancer and has since been dubbed the *tobacco strategy* [1]. More recently, the strategy has been applied to other issues including the link between fossil fuels and acid rain, the link between fossil fuels and global warming, and the link between industrial chemicals and neurological disease [1,2]. The essence of the strategy is to promote the idea that the science underlying certain risks is unresolved, and therefore the implicated technologies should not be regulated. The strategy is typically employed by an interconnected mix of industry interests and ideological opponents of regulation. The target audience is typically a mix of government officials and the general public, and not the scientific community.

As is discussed in more detail below, certain factors suggest the potential for superintelligence to be a focus of risk–profit politicized skepticism. First and foremost, superintelligence could be developed by major corporations with a strong financial incentive to avoid regulation. Second, there already exists a lot of skepticism about superintelligence, which could be exploited for political purposes. Third, as an unprecedented class of technology, it is inherently uncertain, which suggests that superintelligence skepticism may be especially durable, even within apolitical scholarly communities. These and other factors do not guarantee that superintelligence skepticism will be politicized, or that its politicization would follow the same risk–profit patterns as the tobacco strategy. However, these factors are at least suggestive of the possibility.

Superintelligence skepticism may also be politicized in a different way: to protect the reputations and funding of the broader AI field. This form of politicized skepticism is less well-documented than the tobacco strategy, and appears to be less common. However, there are at least hints of it for fields of technology involving both grandiose future predictions and more mundane near-term work. AI is one such field of technology, in which grandiose predictions of superintelligence and other future AI breakthroughs contrast with more modest forms of near-term AI. Another example is nanotechnology, in which grandiose predictions of molecular machines contrast with near-term nanoscale science and technology [3].

The basis of the paper's analysis is twofold. First, the paper draws on the long history of risk–profit politicized skepticism. This history suggests certain general themes that may also apply to superintelligence. Second, the paper examines characteristics of superintelligence development to assesses the prospect of skepticism being used politically in this context. To that end, the paper draws on the current state of affairs in the AI sector, especially for artificial general intelligence, which is a type of AI closely related to superintelligence. The paper further seeks to inform efforts to avoid any potential harmful effects from politicized superintelligence skepticism. The effects would not necessarily be harmful, but the history of risk–profit politicized skepticism suggests that they could be.

This paper contributes to literatures on politicized skepticism and superintelligence governance. Whereas most literature on politicized skepticism (and similar concepts such as denial) is backward-looking, consisting of historical analysis of skepticisms that have already occurred [1,2,4–7], this paper is largely (but not exclusively) forward-looking, consisting of prospective analysis of skepticisms that could occur at some point in the future. Meanwhile, the superintelligence governance literature has looked mainly at institutional regulations to prevent research groups from building dangerous superintelligence and support for research on safety measures [8–11]. This paper contributes to a smaller literature on the role of corporations in superintelligence development [12] and on social and psychological aspects of superintelligence governance [13].

This paper does not intend to take sides on which beliefs about superintelligence are most likely to be correct. Its interest is in the potential political implications of superintelligence skepticism, not in the underlying merits of the skepticism. The sole claim here is that the possibility of politicized superintelligence skepticism is a worthy topic of study. It is worth studying due to: (1) the potential for large consequences if superintelligence is built; and (2) the potential for superintelligence to be an important political phenomenon regardless of whether it is built. Finally, the topic is also of inherent intellectual interest as an exercise in prospective socio-political analysis on a possible future technology.

The paper is organized as follows. Section 2 presents a brief overview of superintelligence concerns and skepticisms. Section 3 further develops the concept of politicized skepticism and surveys the history of risk–profit politicized skepticism, from its roots in tobacco to the present day. Section 4 discusses prospects for politicized superintelligence skepticism. Section 5 discusses opportunities for constructive action. Section 6 concludes.

2. Superintelligence and Its Skeptics

The idea of humans being supplanted by their machines dates to at least the 1863 work of Butler [14]. In 1965, Good presented an early exposition on the topic within the modern field of computer science [15]. Good specifically proposed an "intelligence explosion" in which intelligent machines make successively more intelligent machines until they are much smarter than humans, which would be "the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control" [15] (p. 33). This intelligence explosion is one use of the term *technological singularity*, though the term can also refer to wider forms of radical technological change [16]. The term *superintelligence* refers specifically to AI that is much more intelligent than humans and dates to at least the 1998 work of Bostrom [17]. A related term is *artificial general intelligence*, which is AI capable of reasoning across many intellectual domains. A superintelligent AI is likely to have general intelligence, and the development of artificial general intelligence could be a major precursor to superintelligence. Artificial general intelligence is also an active subfield of AI [18,19].

Superintelligence is notable as a potential technological accomplishment with massive societal implications. The effects of superintelligence could include anything from solving a significant portion of the world's problems (if superintelligence is designed well) to causing the extinction of humans and other species (if it is designed poorly). Much of the interest in superintelligence derives from these high stakes. Superintelligence is also of intellectual interest as perhaps the ultimate accomplishment within the field of AI, sometimes referred to as the "grand dream" of AI [20] (p. 125).

Currently, most AI research is on *narrow AI* that is not oriented towards this grand dream. The focus on narrow AI dates to early struggles in the field to make progress towards general AI or superintelligence. After an initial period of hype fell short, the field went through an "AI winter" marked by diminished interest and more modest expectations [21,22] This prompted a focus on smaller, incremental progress on narrow AI. It should be noted that the term AI winter most commonly refers to a lull in AI in the mid-to-late 1980s and early 1990s. A similar lull occurred in the 1970s, and concerns about a new winter can be found as recently as 2008 [23].

With most of the field focused on narrow AI, artificial general intelligence has persisted only as a small subfield of AI [18]. The AI winter also caused many AI computer scientists to be skeptical of superintelligence, on grounds that superintelligence has turned out to be much more difficult than initially expected, and likewise to be averse to attention to superintelligence, on grounds that such hype could again fall short and induce another AI winter. This is an important historical note because it indicates that superintelligence skepticism has wide salience across the AI computer science community and may already be politicized towards the goal of protecting the reputation of and funding for AI. (More on this below.)

Traces of superintelligence skepticism predate AI winter. Early AI skepticism dates to 1965 work by Dreyfus [24]. Dreyfus [24] critiqued the overall field of AI, with some attention to human-level AI though not to superintelligence. Dreyfus traced this skepticism of machines matching human intelligence to a passage in Descartes' 1637 *Discourse On Method* [25]: "it must be morally impossible that there should exist in any machine a diversity of organs sufficient to enable it to act in all the occurrences of life, in the way in which our reason enables us to act."

In recent years, superintelligence has attracted considerable attention. This has likely been prompted by several factors, including a growing scholarly literature (e.g., [9,19,26–29]), highly publicized remarks by several major science and technology celebrities (e.g., Bill Gates [30], Stephen Hawking [31], and Elon Musk [32]), and breakthroughs in the broader field of AI, which draw attention to AI and may make the prospect of superintelligence seem more plausible (e.g., [33,34]). This attention to superintelligence has likewise prompted some more outspoken skepticism. The following is a brief overview of the debate, including both the arguments of the debate and some biographical information about the debaters. (Biographical details are taken from personal and institutional webpages and are accurate as of the time of this writing, May 2018; they are not necessarily accurate as of the time of the publication of the cited literature.) The biographies can be politically significant because, in public debates, some people's words carry more weight than others'. The examples presented below are intended to be illustrative and at least moderately representative of the arguments made in existing superintelligence skepticism (some additional examples are presented in Section 4). A comprehensive survey of superintelligence skepticism is beyond the scope of this paper.

2.1. Superintelligence Cannot Be Built

Bringsjord [35] argued that superintelligence cannot be built based on reasoning from computational theory. Essentially, the argument is that superintelligence requires a more advanced class of computing, which cannot be produced by humans or existing AI. Bringsjord is Professor of Cognitive Science at Rensselaer Polytechnic University and Director of the Rensselaer AI and Reasoning Lab. Chalmers [36] countered that superintelligence does not necessarily require a more advanced class of computing. Chalmers is University Professor of Philosophy and Neural Science at New York University and co-director of the NYU Center for Mind, Brain, and Consciousness.

McDermott [37] argued that advances in hardware and algorithms may be sufficient to exceed human intelligence, but not to massively exceed it. McDermott is Professor of Computer Science at Yale University. Chalmers [36] countered that, while there may be limits to the potential advances in hardware and software, these limits may not be so restrictive as to preclude superintelligence.

2.2. Superintelligence Is Not Imminent Enough to Merit Attention

Crawford [38] argued that superintelligence is a distraction from issues with existing AI, especially AI that worsens inequalities. Crawford is co-founder and co-director of the AI Now Research Institute at New York University, a Senior Fellow at the NYU Information Law Institute, and a Principal Researcher at Microsoft Research.

Ng argued that superintelligence may be possible, but it is premature to worry about, in particular because it is too different from existing AI systems. Ng memorably likened worrying about superintelligence to worrying about "overpopulation on Mars" [39]. Ng is Vice President and Chief Scientist of Baidu, Co-Chairman and Co-Founder of Coursera, and an Adjunct Professor of Computer Science at Stanford University.

Etzioni [40] argued that superintelligence is unlikely to be built within the next 25 years and is thus not worth current attention. Etzioni is Chief Executive Officer of the Allen Institute for Artificial Intelligence and Professor of Computer Science at University of Washington. Dafoe and Russell [41] countered that superintelligence is worth current attention even if it would take more than 25 years to build. Dafoe is Assistant Professor of Political Science at Yale University and Co-Director of the Governance of AI Program at the University of Oxford. Russell is Professor of Computer Science at University of California, Berkeley. (An alternative counter is that some measures to improve AI outcomes apply to both near-term AI and superintelligence, and thus it is not essential to debate which of the two types of AI should be prioritized [42].)

2.3. Superintelligence Would (Probably) Not Be Catastrophic

Goertzel [43] argued that superintelligence could be built and is worth paying attention to, but also that superintelligence is less likely to result in catastrophe than is sometimes suggested. Specifically, Goertzel argued that it may be somewhat difficult, but very difficult, to build superintelligence with values that are considered desirable, and that the human builders of superintelligence would have good opportunities to check that the superintelligence has the right values. Goertzel is the lead for the OpenCog and SingularityNET projects for developing artificial general intelligence. Goertzel [43] wrote in response to Bostrom [28], who suggested that, if built, superintelligence is likely to result in catastrophe. Bostrom is Professor of Applied Ethics at University of Oxford and Director of the Oxford Future of Humanity Institute. (For a more detailed analysis of this debate, see [44].)

Views similar to Goertzel [43] were also presented by Bieger et al. [45], in particular that the AI that is the precursor to superintelligence could be trained by its human developers to have safe and desirable values. Co-authors Bieger and Thórisson are Ph.D. student and Professor of Computer Science at Reykjavik University; co-author Wang is Associate Professor of Computer and Information Sciences at Temple University.

Searle [46] argued that superintelligence is unlikely to be catastrophic, because it would be an unconscious machine incapable of deciding for itself to attack humanity, and thus humans would need to explicitly program it to cause harm. Searle is Professor Emeritus of the Philosophy of Mind and Language at the University of California, Berkeley. Searle [46] wrote in response to Bostrom [28], who arqued that superintelligence could be dangerous to humans regardless of whether it is conscious.

3. Skepticism as a Political Tool

3.1. The Concept of Politicized Skepticism

There is a sense in which any stated skepticism can be political, insofar as it seeks to achieve certain desired changes within a group. Even the most honest intellectual skepticism can be said to achieve the political aim of advancing a certain form of intellectual inquiry. However, this paper uses the term "politicized skepticism" more narrowly to refer to skepticism with other, non-intellectual aims.

Even with this narrower conception, the distinction between intellectual and politicized skepticism can in practice be blurry. The same skeptical remark can serve both intellectual and (non-intellectual) political aims. People can also have intellectual skepticism that is shaped, perhaps subconsciously, by political factors, as well as politicized skepticism that is rooted in honest intellectual beliefs. For example, intellectuals (academics and the like) commonly have both intellectual and non-intellectual aims, the latter including advancing their careers or making the world a better place per whatever notion of "better" they subscribe to. This can be significant for superintelligence skepticism aimed at protecting the reputations and funding of AI researchers.

It should be stressed that the entanglement of intellectual inquiry and (non-intellectual) political aims does not destroy the merits of intellectual inquiry. This is important to bear in mind at a time when trust in science and other forms of expertise is dangerously low [47,48]. Scholarship can be a social and political process, but, when performed well, it can nonetheless deliver important insights about the world. For all people, scholars included, improving one's understanding of the world takes mental effort, especially when one is predisposed to believe otherwise. Unfortunately, many people are not inclined to make the effort, and other people are making efforts to manipulate ideas for their own aims. An understanding of politicized skepticism is essential for addressing major issues in this rather less-than-ideal epistemic era.

Much of this paper is focused on risk–profit politicized skepticism, i.e., skepticism about concerns about risky and profitable technologies and products. Risk–profit politicized skepticism is a major social force, as discussed throughout this paper, although it is not the only form of politicized skepticism. Other forms include politicized skepticism by concerned citizens, such as skepticism about scientific claims that vaccines or nuclear power plants are safe; by religious activists and institutions, expressing skepticism about claims that humans evolved from other species; by politicians and governments, expressing skepticism about events that cast them in an unfavorable light; and by intellectuals as discussed above. Thus, while this paper largely focuses on skepticism aimed at casting doubt about concerns about risky and profitable technologies and products, it should be understood that this is not the only type of politicized skepticism.

3.2. Tobacco Roots

As mentioned above, risk–profit politicized skepticism traces to 1950s debates on the link between tobacco and cancer. Specifically, in 1954, the tobacco industry formed the Tobacco Industry Research Committee, an "effort to foster the impression of debate, primarily by promoting the work of scientists whose views might be useful to the industry" [1] (p. 17). The committee was led by C. C. Little, who was a decorated genetics researcher and past president of the University of Michigan, as well as a eugenics advocate who believed cancer was due to genetic weakness and not to smoking.

In the 1950s, there was substantial evidence linking tobacco to cancer, but it was not as conclusive of a link as is now available. The tobacco industry exploited this uncertainty in public discussions of the issue. It succeeded in getting major media to often present the issue as a debate between scientists who agreed vs. disagreed in the tobacco–cancer link. Among the media figures to do this was the acclaimed journalist Edward Murrow, himself a smoker who, in tragic irony, later died from lung cancer. Oreskes and Conway speculated that, "Perhaps, being a smoker, he was reluctant to admit that his daily habit was deadly and reassured to hear that the allegations were unproven" [1] (pp. 19–20).

Over subsequent decades, the tobacco industry continued to fund work that questioned the tobacco–cancer link, enabling it to dodge lawsuits and regulations. Then, in 1999, the United States Department of Justice filed a lawsuit against nine tobacco companies and two tobacco trade organizations (United States v. Philip Morris). The US argued that the tobacco industry conspired over several decades to deceive the public, in violation of the Racketeer Influenced and Corrupt Organizations (RICO) Act, which covers organized crime. In 2006, the US District Court for the District of Columbia found the tobacco industry guilty, upheld unanimously in 2009 by the US Court of Appeals. This ruling and other measures have helped to protect people from lung cancer, but many more could have also avoided lung cancer were it not for the tobacco industry's politicized skepticism.

3.3. The Character and Methods of Risk–Profit Politicized Skepticism

The tobacco case provided a blueprint for risk–profit politicized skepticism that has since been used for other issues. Writing in the context of politicized environmental skepticism, Jacques et al. [4] (pp. 353–354) listed four overarching themes: (1) rejection of scientific findings of environmental problems; (2) de-prioritization of environmental problems relative to other issues; (3) rejection of government regulation of corporations and corporate liability; and (4) portrayal of environmentalism as a threat to progress and development. The net effect is to reduce interest in government regulation of corporate activities that may pose harms to society.

The two primary motivations of risk–profit politicized skepticism are the protection of corporate profits and the advancement of anti-regulatory political ideology. The protection of profits is straightforward: from the corporation's financial perspective, the investment in politicized skepticism can bring a substantial return. The anti-regulatory ideology is only slightly subtler. Risk–profit politicized skepticism is often associated with pro-capitalist, anti-socialist, and anti-communist politics. For example, some political skeptics liken environmentalists to watermelons: "green on the outside, red on the inside" [1] (p. 248), while one feared that the Earth Summit was a socialist plot to establish a "World Government with central planning by the United Nations" [1] (p. 252). For these people, politicized skepticism is a way to counter discourses that could harm their political agenda.

Notably, both the financial and the ideological motivations are not inherently about science. Instead, the science is manipulated towards other ends. This indicates that the skepticism is primarily political and not intellectual. It may still be intellectually honest in the sense that the people stating the skepticism are actually skeptical. That would be consistent with author Upton Sinclair's saying that "It is difficult to get a man to understand something when his salary depends upon his not understanding it." The skepticism may nonetheless violate that essential intellectual virtue of letting conclusions follow from analysis, and not the other way around. For risk–profit politicized skepticism, the desired conclusion is typically the avoidance of government regulation of corporate activity, and the skepticism is crafted accordingly.

To achieve this end, the skeptics will often engage in tactics that clearly go beyond honest intellectual skepticism and ordinary intellectual exchange. For example, ExxonMobil has been found to express extensive skepticism about climate change in its public communications (such as newspaper advertisements), but much less skepticism in its internal communications and peer-reviewed publications [7]. This finding suggests that ExxonMobil was aware of the risks of climate change and misled the public about the risks. ExxonMobil reportedly used its peer-reviewed publications for "the credentials required to speak with authority in this area", including in its conversations with government officials [7] (p. 15), even though these communications may have presented climate change risk differently than the peer-reviewed publications did. (As an aside, it may be noted that the ExxonMobil study [7], published in 2017, has already attracted a skeptic critique by Stirling [49]. Stirling is Communications Manager of the Canadian nonprofit Friends of Science. Both Stirling and Friends of Science are frequent climate change skeptics [50].)

While the skeptics do not publicly confess dishonesty, there are reports that some of them have privately done so. For example, Marshall [51] (p. 180) described five energy corporation presidents who believed that climate change was a problem and "admitted, off the record, that the competitive environment forced them to suppress the truth about climate change" to avoid government regulations. Similarly, US Senator Sheldon Whitehouse, an advocate of climate policy to reduce greenhouse gas emissions, reported that some of his colleagues publicly oppose climate policy but privately support it, with one even saying "Let's keep talking—but don't tell my staff. Nobody else can know" [52] (p. 176). Needless to say, any instance in which skepticism is professed by someone who is not actually skeptical is a clear break from the intellectual skepticism of ordinary scholarly inquiry.

One particularly distasteful tactic is to target individual scientists, seeking to discredit their work or even intimidate them. For example, Philippe Grandjean, a distinguished environmental health researcher, reported that the tuna industry once waged a $25 million advertising campaign criticizing work by himself and others who have documented links between tuna, mercury, and neurological disease. Grandjean noted that $25 million is a small sum for the tuna industry but more than the entire sum of grant funding he received for mercury research over his career, indicating a highly uneven financial playing field [2] (pp. 119–120). In another example, climate scientists accused a climate skeptic of bullying and intimidation and reported receiving "a torrent of abusive and threatening e-mails after being featured on" the skeptic's blog, which calls for climate scientists "to be publicly flogged" [51] (p. 151).

Much of the work, however, is far subtler than this. Often, it involves placing select individuals in conferences, committees, or hearings, where they can ensure that the skeptical message is heard in the right places. For example, Grandjean [2] (p. 129) recounted a conference sponsored by the Electric Power Research Institute, which gave disproportionate floor time to research questioning the health effects of mercury. In another episode, the tobacco industry hired a recently retired World Health Organization committee chair to "volunteer" as an advisor to the same committee, which then concluded to not restrict use of a tobacco pesticide [2] (p. 125).

Another common tactic is to use outside organizations as the public face of the messaging. This tactic is accused of conveying the impression that the skepticism is done in the interest of the public and not of private industry. Grandjean [2] (p. 121) wrote that "organizations, such as the Center for Science and Public Policy the Center for Indoor Air Research or the Citizens for Fire Safety Institute, may sound like neutral and honest establishments, but they turned out to be 'front groups' for financial interests." Often, the work is done by think tanks. Jacques et al. [4] found that over 90% of books exhibiting environmental skepticism are linked to conservative think tanks, and 90% of conservative think tanks are active in environmental skepticism. This finding is consistent with recent emphasis in US conservatism on unregulated markets. (Earlier strands of US conservatism were more supportive of environmental protection, such as the pioneering American conservative Russell Kirk, who wrote that "There is nothing more conservative than conservation" [53].)

3.4. The Effectiveness of Politicized Skepticism

Several broader phenomena help make politicized skepticism so potent, especially for risk–profit politicized skepticism. One is the enormous amounts of corporate money at stake with certain government regulations. When corporations use even a tiny fraction of this for politicized skepticism, it can easily dwarf other efforts. Similarly, US campaign finance laws are highly permissive.

Whitehouse [52] traced the decline in bipartisan Congressional support for climate change policy to the Supreme Court's 2010 *Citizens United* ruling, which allows unlimited corporate spending in elections. However, even without election spending, corporate assets tilt the playing field substantially in the skeptics' favor.

Another important factor is the common journalistic norm of balance, in which journalists seek to present "both sides" of an issue. This can put partisan voices on equal footing with independent science, as seen in early media coverage of tobacco. It can also amplify a small minority of dissenting voices, seen more recently in media coverage of climate change. Whereas the scientific community has overwhelming consensus that climate change is happening, that it is caused primarily by human activity, and that the effects will be mainly harmful, public media features climate change skepticism much more than its scientific salience would suggest [54]. (For an overview of the scientific issues related to climate change skepticism, see [55]; for documentation of the scientific consensus, see [56].)

A third factor is the tendency of scientists to be cautious with respect to uncertainty. Scientists often aspire to avoid stating anything incorrect and to focus on what can be rigorously established instead of discussing more speculative possibilities. Scientists will also often highlight remaining uncertainties even when basic trends are clear. "More research is needed" is likely the most ubiquitous conclusion of any scientific research. This tendency makes it easier for other parties to make the state of the science appear less certain than it actually is. Speaking to this point in a report on climate change and national security, former US Army Chief of Staff Gordon Sullivan states "We seem to be standing by and, frankly, asking for perfectness in science . . . We never have 100 percent certainty. We never have it. If you wait until you have 100 percent certainty, something bad is going to happen on the battlefield" [57] (p. 10).

A fourth factor is the standard, found in some (but not all) policy contexts, of requiring robust evidence of harm before pursuing regulation. In other words, the burden of proof is on those who wish to regulate, and the potentially harmful product is presumed innocent until proven guilty. Grandjean [2] cited this as the most important factor preventing the regulation of toxic chemicals in the US. Such a protocol makes regulation very difficult, especially for complex risks that resist precise characterization. In these policy contexts, the amplification of uncertainty can be particularly impactful.

To sum up, risk–profit politicized skepticism is a longstanding and significant tool used to promote certain political goals. It has been used heavily by corporations seeking to protect profits and people with anti-regulatory ideologies, and it has proven to be a powerful tool. In at least one case, the skeptics were found guilty in a court of law of conspiracy to deceive the public. The skeptics use a range of tactics that deviate from standard intellectual practice, and they exploit several broader societal phenomena that make the skepticism more potent.

4. Politicized Superintelligence Skepticism

4.1. Is Superintelligence Skepticism Already Politicized?

At this time, there does not appear to be any superintelligence skepticism that has been politicized to the extent that has occurred for other issues such as tobacco–cancer and fossil fuels–global warming. Superintelligence skeptics are not running ad campaigns or other major dollar operations. For the most part, they are not attacking the scholars who express concern about superintelligence. Much of the discussion appears in peer-reviewed journals, and has the tone of constructive intellectual discourse. An exception that proves the rule is Etzioni [40], who included a quotation comparing Nick Bostrom (who is concerned about superintelligence) to Donald Trump. In a postscript on the matter, Etzioni [40] wrote that "we should refrain from ad hominem attacks. Here, I have to offer an apology". In contrast, the character attacks of the most heated politicized skepticism are made without apology.

However, there are already at least some hints of politicized superintelligence skepticism. Perhaps the most significant comes from AI academics downplaying hype to protect their field's reputation and funding. The early field of AI made some rather grandiose predictions, which soon fell flat, fueling criticisms as early as 1965 [24]. Some of these criticisms prompted major funding cuts, such as the 1973 Lighthill report [58], which prompted the British Science Research Council to slash its support for AI. Similarly, Menzies [59] described AI as going through a "peak of inflated expectations" in the 1980s followed by a "trough of disillusionment" in the late 1980s and early 1990s. Most recently, writing in 2018, Bentley [60] (p. 11) derided beliefs about superintelligence and instead urges: "Do not be fearful of AI—marvel at the persistence and skill of those human specialists who are dedicating their lives to help create it. And appreciate that AI is helping to improve our lives every day." (For criticism of Bentley [60], see [61].) This suggests that some superintelligence skepticism may serve the political goal of protecting the broader field of AI.

Superintelligence skepticism that is aimed at protecting the field of AI may be less of a factor during the current period of intense interest in AI. At least for now, the field of AI does not need to defend its value—its value is rather obvious, and AI researchers are not lacking for job security. Importantly, the current AI boom is largely based on actual accomplishments, not hype. Therefore, while today's AI researchers may view superintelligence as a distraction, they are less likely to view it as a threat to their livelihood. However, some may nonetheless view superintelligence in this way, especially those who have been in the field long enough to witness previous boom-and-bust cycles. Likewise, the present situation could change if the current AI boom eventually cycles into another bust—another winter. Despite the success of current AI, there are arguments that it is fundamentally limited [62]. The prospect of a new AI winter could be a significant factor in politicized superintelligence skepticism.

A different type of example comes from public intellectuals who profess superintelligence skepticism based on questionable reasoning. A notable case of this is the psychologist and public intellectual Steven Pinker. Pinker recently articulated a superintelligence skepticism that some observers have likened to politicized climate skepticism [63,64]. Pinker does resemble some notable political skeptics: a senior scholar with an academic background in an unrelated topic who is able to use his (and it is typically a *he*) platform to advance his skeptical views. Additionally, a close analysis of Pinker's comments on superintelligence finds them to be flawed and poorly informed by existing research [65]. Pinker's superintelligence skepticism appears to be advancing a broader narrative of human progress, and may be making the intellectual sin of putting this conclusion before the analysis of superintelligence. However, his particular motivations are, to the present author's knowledge, not documented (It would be especially ironic for Pinker to politicize skepticism based on flawed intellectual reasoning, since he otherwise preaches a message intellectual virtue).

A third type of example of potential politicized superintelligence skepticism comes from the corporate sector. Several people in leadership positions at technology corporations have expressed superintelligence skepticism, including Eric Schmidt (Executive Chairman of Alphabet, the parent company of Google) [66] and Mark Zuckerberg (CEO of Facebook) [67]. Since this skepticism comes the corporate sector, it has some resemblance to risk–profit politicized skepticism and may likewise have the most potential to shape public discourse and policy. One observer postulated that Zuckerberg professes superintelligence skepticism to project the idea that "software is always friendly and tame" and avoid the idea "that computers are intrinsically risky", the latter of which "has potentially dire consequences for Zuckerberg's business and personal future" [67]. While this may just be conjecture, it does come at a time in which Facebook is under considerable public pressure for its role in propagating fake news and influencing elections, which, although unrelated to superintelligence, nonetheless provides an antiregulatory motivation to downplay risks associated with computers.

To summarize, there may already be some politicized superintelligence skepticism, coming from AI academics seeking to protect their field, public intellectuals seeking to advance a certain narrative about the world, and corporate leaders seeking to avoid regulation. However, it is not clear how much superintelligence skepticism is already politicized, and there are indications that it may be limited, especially compared to what has occurred for other issues. On the other hand, superintelligence is a relatively new public issue (with a longer history in academia), so perhaps its politicization is just beginning.

Finally, it is worth noting that while superintelligence has not been politicized to the extent that climate change has, there is at least one instance of superintelligence being cited in the context of climate skepticism. Cass [68,69] cited the prospect of superintelligence as a reason to not be concerned about climate change. A counter to this argument is that, even if superintelligence is a larger risk, addressing climate change can still reduce the overall risk faced by humanity. Superintelligence could also be a solution to climate change, and thus may be worth building despite the risks it poses. At the same time, if climate change has been addressed independently, then this reduces the need to take risks in building superintelligence [70].

4.2. Prospects for Politicized Superintelligence Skepticism

Will superintelligence skepticism be (further) politicized? Noting the close historical association between politicized skepticism and corporate profits—at least for risk–profit politicized skepticism—an important question is whether superintelligence could prompt profit-threatening regulations. AI is now being developed by some of the largest corporations in the world. Furthermore, a recent survey found artificial general intelligence projects at several large corporations, including Baidu, Facebook, Google, Microsoft, Tencent, and Uber [19]. These corporations have the assets to conduct politicized skepticism that is every bit as large as that of the tobacco, fossil fuel, and industrial chemicals industries.

It should be noted that the artificial general intelligence projects at these corporations were not found to indicate substantial skepticism. Indeed, some of them are outspoken in concern about superintelligence. Moreover, out of 45 artificial general intelligence projects surveyed, only two were found to be dismissive of concerns about the risks posed by the technology [19]. However, even if the AI projects themselves do not exhibit skepticism, the corporations that host them still could. Such a scenario would be comparable to that of ExxonMobil, whose scientists confirmed the science of climate change even while corporate publicity campaigns professed skepticism [7].

The history shows that risk–profit politicized skepticism is not inherent to corporate activity—it is generally only found when profits are at stake. The preponderance of corporate research on artificial general intelligence suggests at least a degree of profitability, but, at this time, it is unclear how profitable it will be. If it is profitable, then corporations are likely to become highly motivated to protect it against outside restrictions. This is an important factor to monitor as the technology progresses.

In public corporations, the pressure to maximize shareholder returns can motivate risk–profit politicized skepticism. However, this may be less of a factor for some corporations in the AI sector. In particular, voting shares constituting a majority of voting power at both Facebook and Alphabet (the parent company of Google) are controlled by the companies' founders: Mark Zuckerberg at Facebook [71] and Larry Page and Sergey Brin at Alphabet [72]. Given their majority stakes, the founders may be able to resist shareholder pressure for politicized skepticism, although it is not certain that they would, especially since leadership at both companies already display superintelligence skepticism.

Another factor is the political ideologies of those involved in superintelligence. As discussed above, risk–profit politicized skepticism of other issues is commonly driven by people with pro-capitalist, anti-socialist, and anti-communist political ideologies. Superintelligence skepticism may be more likely to be politicized by people with similar ideologies. Some insight into this matter can be obtained from a recent survey of 600 technology entrepreneurs [73], which is a highly relevant demographic. The study finds that, contrary to some conventional wisdom, this demographic tends not to hold libertarian ideologies. Instead, technology entrepreneurs tend to hold views consistent with American liberalism, but with one important exception: technology entrepreneurs tend to oppose government regulation. This finding suggests some prospect for politicizing superintelligence skepticism, although perhaps not as much as may exist in other industries.

Further insight can be found from the current political activities of AI corporations. In the US, the corporations' employees donate mainly to the Democratic Party, which is the predominant party of American liberalism and is more pro-regulation. However, the corporations themselves have recently shifted donations to the Republican Party, which is the predominant party of American conservatism and is more anti-regulation. Edsall [74] proposed that this divergence between employees and employers is rooted in corporations' pursuit of financial self-interest. A potential implication of this is that, even if the individuals who develop AI oppose risk–profit politicized skepticism, the corporations that they work for may support it. Additionally, the corporations have recently been accused of using their assets to influence academic and think tank research on regulations that the corporations could face [75,76], although at least some of the accusations have been disputed [77]. While the veracity of these accusations is beyond the scope of this paper, they are at least suggestive of the potential for these corporations to politicize superintelligence skepticism.

AI corporations would not necessarily politicize superintelligence skepticism, even if profits may be at stake. Alternatively, they could express concern about superintelligence to portray themselves as responsible actors and likewise avoid regulation. This would be analogous to the strategy of "greenwashing" employed by companies seeking to bolster their reputation for environmental stewardship [78]. Indeed, there have already been some expressions of concern about superintelligence by AI technologists, and likewise some suspicion that the stated concern has this sort of ulterior motive [79].

To the extent that corporations do politicize superintelligence skepticism, they are likely to mainly emphasize doubt about the risks of superintelligence. Insofar as superintelligence could be beneficial, corporations may promote this, just as they promote the benefits of fossil fuels (for transportation, heating, etc.) and other risky products. Or, AI corporations may promote the benefits of their own safety design and sow doubt about the safety of their rivals' designs, analogous to the marketing of products whose riskiness can vary from company to company, such as automobiles. Alternatively, AI corporations may seek to sow doubt about the possibility of superintelligence, calculating that this would be their best play for avoiding regulation. As with politicized skepticism about other technologies and products, there is no one standard formula that every company always adopts.

For their part, academic superintelligence skeptics may be more likely to emphasize doubt about the mere possibility of superintelligence, regardless of whether it would be beneficial or harmful, due to reputational concerns. Or, they could focus skepticism on the risks, for similar reasons as corporations: academic research can also be regulated, and researchers do not always welcome this. Of course, there are also academics who do not exhibit superintelligence skepticism. Again, there is no one standard formula.

4.3. Potential Effectiveness of Politicized Superintelligence Skepticism

If superintelligence skepticism is politicized, several factors point to it being highly effective, even more so than for the other issues in which skepticism has been politicized.

First, some of the experts best positioned to resolve the debate are also deeply implicated in it. To the extent that superintelligence is a risk, the risk is driven by the computer scientists who would build superintelligence. These individuals have intimate knowledge of the technology and thus have an essential voice in the public debate (though not the only essential voice). This is distinct from issues such as tobacco or climate change, in which the risk is mainly assessed by outside experts. It would be as if the effect of tobacco on cancer was studied by the agronomists who cultivate tobacco crops, or if the science of climate change was studied by the geologists who map deposits of fossil fuels. With superintelligence, a substantial portion of the relevant experts have a direct incentive to avoid any restrictions on the technology, as do their employers. This could create a deep and enduring pool of highly persuasive skeptics.

Second, superintelligence skepticism has deep roots in the mainstream AI computer science community. As noted above, this dates to the days of AI winter. Thus, skeptics may be abundant even where they are not funded by industry. Indeed, most of the skeptics described above do not appear to be speaking out of any industry ties, and thus would not have an industry conflict of interest. They could still have a conflict of interest from their desire in protect the reputation of their field, but this is a subtler matter. Insofar as they are perceived to not have a conflict of interest, they could be especially persuasive. Furthermore, even if their skepticism is honest and not intended for any political purposes, it could be used by others in dishonest and political ways.

Third, superintelligence is a topic for which the uncertainty is inherently difficult to resolve. It is a hypothetical future technology that is qualitatively different from anything that currently exists. Furthermore, there is concern that its mere existence could be catastrophic, which could preclude certain forms of safety testing. It is thus a risk that defies normal scientific study. In this regard, it is similar to climate change: moderate climate change can already be observed, as can moderate forms of AI, but the potentially catastrophic forms have not yet materialized and possibly never will. However, climate projections can rely on some relatively simple physics—at its core, climate change largely reduces to basic physical chemistry and thermodynamics. (The physical chemistry covers the nature of greenhouse gasses, which are more transparent to some wavelengths of electromagnetic radiation than to others. The thermodynamics covers the heat transfer expected from greenhouse gas buildup. Both effects can be demonstrated in simple laboratory experiments. Climate change also involves indirect feedback effects on much of the Earth system, including clouds, ice, oceans, and ecosystems, which are often more complex and difficult to resolve and contribute to ongoing scientific uncertainty.) In contrast, AI projections must rely on notions of intelligence, which is not so simple at all. For this reason, it is less likely that scholarly communities will converge on any consensus position on superintelligence in the way that they have on other risks such as climate change.

Fourth, some corporations that could develop superintelligence may be uniquely well positioned to influence public opinion. The corporations currently involved in artificial general intelligence research include some corporations that also play major roles in public media. As a leading social media platform, Facebook in particular has been found to be especially consequential for public opinion [80]. Corporations that serve as information gateways, such as Baidu, Google, and Microsoft, also have unusual potential for influence. These corporations have opportunities to shape public opinion in ways that the tobacco, fossil fuel, and industrial chemicals industries cannot. While the AI corporations would not necessarily exploit these opportunities, it is an important factor to track.

In summary, while it remains to be seen whether superintelligence skepticism will be politicized, there are some reasons for believing it will be, and that superintelligence would be an especially potent case of politicized skepticism.

5. Opportunities for Constructive Action

Politicized superintelligence skepticism would not necessarily be harmful. As far as this paper is concerned, it is possible that, for superintelligence, skepticism is the correct view, meaning that superintelligence may not be built, may not be dangerous, or may not merit certain forms of imminent attention. (The paper of course assumes that superintelligence is worth some imminent attention, or otherwise it would not have been written.) It is also possible that, even if superintelligence is a major risk, government regulations could nonetheless be counterproductive, and politicized skepticism could help avoid that. That said, the history of politicized skepticism (especially risk–profit politicized skepticism) shows a tendency for harm, which suggests that politicized superintelligence skepticism could be harmful as well.

With this in mind, one basic opportunity is to raise awareness about politicized skepticism within communities that discuss superintelligence. Superintelligence skeptics who are motivated by honest intellectual norms may not wish for their skepticism to be used politically. They can likewise be cautious about how to engage with potential political skeptics, such as by avoiding certain speaking opportunities in which their remarks would be used as a political tool instead of as a constructive intellectual contribution. Additionally, all people involved in superintelligence debates can insist on basic intellectual standards, above all by putting analysis before conclusions and not the other way around. These are the sorts of things that an awareness of politicized skepticism can help with.

Another opportunity is to redouble efforts to build scientific consensus on superintelligence, and then to draw attention to it. Currently, there is no consensus. As noted above, superintelligence is an inherently uncertain topic and difficult to build consensus on. However, with some effort, it should be possible to at least make progress towards consensus. Of course, scientific consensus does not preclude politicized skepticism—ongoing climate skepticism attests to this. However, it can at least dampen the politicized skepticism. Indeed, recent research has found that the perception of scientific consensus increases acceptance of the underlying science [81].

A third opportunity is to engage with AI corporations to encourage them to avoid politicizing skepticism about superintelligence or other forms of AI. Politicized skepticism is not inevitable, and while corporate leaders may sometimes feel as though they have no choice, there may nonetheless be options. Furthermore, the options may be especially effective at this early stage in superintelligence research, in which corporations may have not yet established internal policy or practices.

A fourth opportunity is to follow best practices in debunking misinformation in the event that superintelligence skepticism is politicized. There is a substantial literature on the psychology of debunking [81–83]. A debunking handbook written for a general readership [82] recommends: (1) focusing on the correct information to avoid cognitively reinforcing the false information; (2) preceding any discussion of the false information with a warning that it is false; and (3) when debunking false information, also give the correct information so that people are not left with a gap in their understanding of the topic. The handbook further cautions against using the *information deficit model* of human cognition, which proposes that mistaken beliefs can be corrected simply by providing the correct information. The information deficit model is widely used in science communication, but it has been repeatedly found to work poorly, especially in situations of contested science. This sort of advice could be helpful to efforts to counter superintelligence misinformation.

Finally, the entire AI community should insist that policy be made based on an honest and balanced read of the current state of knowledge. Burden of proof requirements should not be abused for private gain. As with climate change and other global risks, the world cannot afford to prove that superintelligence would be catastrophic. By the time uncertainty is eliminated, it could be too late.

6. Conclusions

Some people believe that superintelligence could be a highly consequential technology, potentially even a transformative event in the course of human history, with either profoundly beneficial or extremely catastrophic effects. Insofar as this belief is plausible, superintelligence may be worth careful advance consideration, to ensure that the technology is handled successfully. Importantly, this advance attention should include social science and policy analysis, and not just computer science. Furthermore, even if belief in superintelligence is mistaken, it can nonetheless be significant as a social and political phenomenon. This is another reason for social science and policy analysis. This paper is a contribution to the social science and policy analysis of superintelligence. Furthermore, despite the unprecedented nature of superintelligence, this paper shows that there are important historical and contemporary analogs that can shed light on the issue. Much of what could occur for the development of superintelligence has already occurred for other technologies. Politicized skepticism is one example of this.

One topic not covered in this paper is the prospect of beliefs that superintelligence will occur and/or will be harmful to be politicized. Such a phenomenon could be analogous to, for example, belief in large medical harms from nuclear power, or, phrased differently, skepticism about claims that nuclear power plants are medically safe. The scientific literature on nuclear power finds medical harms to be substantially lower than is commonly believed [84]. Overstated concern (or "alarmism") about nuclear power can likewise be harmful, for example by increasing use of fossil fuels. Similarly, the fossil fuel industry could politicize this belief for its own benefit. By the same logic, belief in superintelligence could also be politicized. This prospect is left for future research, although much of this paper's analysis may be applicable.

Perhaps the most important lesson of this paper is that the development of superintelligence could be a contentious political process. It could involve aggressive efforts by powerful actors—efforts that not only are inconsistent with basic intellectual ideals, but that also actively subvert those ideals for narrow, self-interested gain. This poses a fundamental challenge to those who seek to advance a constructive study of superintelligence.

Funding: This research received no external funding.

Acknowledgments: Tony Barrett, Phil Torres, Olle Häggström, Maurizio Tinnirello, Matthijs Maas, Roman Yampolskiy, and participants in a seminar hosted by the Center for Human-Compatible AI at UC Berkeley provided helpful feedback on an earlier version of this paper. All remaining errors are the author's alone. The views expressed in this paper are the author's and not necessarily the views of the Global Catastrophic Risk Institute.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Oreskes, N.; Conway, E.M. *Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming*; Bloomsbury: New York, NY, USA, 2010.
2. Grandjean, P. *Only One Chance: How Environmental Pollution Impairs Brain Development—And How to Protect the Brains of the Next Generation*; Oxford University Press: Oxford, UK, 2013.
3. Selin, C. Expectations and the emergence of nanotechnology. *Sci. Technol. Hum. Values* **2007**, *32*, 196–220. [CrossRef]
4. Jacques, P.J.; Dunlap, R.E.; Freeman, M. The organisation of denial: Conservative think tanks and environmental skepticism. *Environ. Politics* **2008**, *17*, 349–385. [CrossRef]
5. Lewandowsky, S.; Oberauer, K. Motivated rejection of science. *Curr. Dir. Psychol. Sci.* **2016**, *25*, 217–222. [CrossRef]
6. Lewandowsky, S.; Mann, M.E.; Brown, N.J.; Friedman, H. Science and the public: Debate, denial, and skepticism. *J. Soc. Polit. Psychol.* **2016**, *4*, 537–553. [CrossRef]
7. Supran, G.; Oreskes, N. Assessing ExxonMobil's climate change communications (1977–2014). *Environ. Res. Lett.* **2017**, *12*, 084019. [CrossRef]
8. McGinnis, J.O. Accelerating Ai. *Northwest. Univ. Law Rev.* **2010**, *104*, 366–381. [CrossRef]
9. Sotala, K.; Yampolskiy, R.V. Responses to catastrophic AGI risk: A survey. *Phys. Scr.* **2014**, *90*, 018001. [CrossRef]
10. Wilson, G. Minimizing global catastrophic and existential risks from emerging technologies through international law. *VA Environ. Law J.* **2013**, *31*, 307–364.
11. Yampolskiy, R.; Fox, J. Safety engineering for artificial general intelligence. *Topoi* **2013**, *32*, 217–226. [CrossRef]
12. Goertzel, B. The Corporatization of AI Is a Major Threat to Humanity. *H+ Magazine*, 21 July 2017.
13. Baum, S.D. On the promotion of safe and socially beneficial artificial intelligence. *AI Soc.* **2017**, *32*, 543–551. [CrossRef]
14. Butler, S. Darwin among the Machines. *The Press*, 13 June 1863.
15. Good, I.J. Speculations concerning the first ultraintelligent machine. *Adv. Comput.* **1965**, *6*, 31–88.
16. Sandberg, A. An overview of models of technological singularity. In *The Transhumanist Reader: Classical and Contemporary Essays on the Science, Technology, and Philosophy of the Human Future*; More, M., Vita-More, N., Eds.; Wiley: New York, NY, USA, 2010; pp. 376–394.
17. Bostrom, N. How Long before Superintelligence? 1998. Available online: https://nickbostrom.com/superintelligence.html (accessed on 18 August 2018).
18. Goertzel, B. Artificial general intelligence: Concept, state of the art, and future prospects. *J. Artif. Gen. Intell.* **2014**, *5*, 1–48. [CrossRef]
19. Baum, S.D. A Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy. Global Catastrophic Risk Institute Working Paper 17-1, 2017. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3070741 (accessed on 18 August 2018).

20. Legg, S. Machine Super Intelligence. Ph.D. Thesis, University of Lugano, Lugano, Switzerland, 2008.
21. Crevier, D. *AI: The Tumultuous History of the Search for Artificial Intelligence*; Basic Books: New York, NY, USA, 1993.
22. McCorduck, P. *Machines Who Think: 25th Anniversary Edition*; A.K. Peters: Natick, MA, USA, 2004.
23. Hendler, J. Avoiding another AI winter. *IEEE Intell. Syst.* **2008**, *23*, 2–4. [CrossRef]
24. Dreyfus, H. Alchemy and AI. RAND Corporation Document P-3244, 1965. Available online: https://www.rand.org/pubs/papers/P3244.html (accessed on 18 August 2018).
25. Descartes, R. *A Discourse on Method*; Project Gutenberg eBook, 1637. Available online: http://www.gutenberg.org/files/59/59-h/59-h.htm (accessed on 18 August 2018).
26. Chalmers, D.J. The singularity: A philosophical analysis. *J. Conscious. Stud.* **2010**, *17*, 7–65.
27. Eden, A.H.; Moor, J.H.; Soraker, J.H.; Steinhart, E. (Eds.) *Singularity Hypotheses: A Scientific and Philosophical Assessment*; Springer: Berlin, Germany, 2013.
28. Bostrom, N. *Superintelligence: Paths, Dangers, Strategies*; Oxford University Press: Oxford, UK, 2014.
29. Callaghan, V.; Miller, J.; Yampolskiy, R.; Armstrong, S. (Eds.) *The Technological Singularity: Managing the Journey*; Springer: Berlin, Germany, 2017.
30. Rawlinson, K. Microsoft's Bill Gates Insists AI Is a Threat. *BBC*, 29 January 2015.
31. Cellan-Jones, R. Stephen Hawking Warns Artificial Intelligence Could End Mankind. *BBC*, 2 December 2014.
32. Dowd, M. Elon Musk's Billion-Dollar Crusade to Stop the A.I. Apocalypse. *Vanity Fair*, 26 March 2017.
33. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
34. Silver, D.; Huang, A.; Maddison, C.J.; Guez, A.; Sifre, L.; Van Den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; et al. Mastering the game of Go with deep neural networks and tree search. *Nature* **2016**, *529*, 484–489. [CrossRef] [PubMed]
35. Bringsjord, S. Belief in the singularity is logically brittle. *J. Conscious. Stud.* **2012**, *19*, 14–20.
36. Chalmers, D. The Singularity: A reply. *J. Conscious. Stud.* **2012**, *19*, 141–167.
37. McDermott, D. Response to the singularity by David Chalmers. *J. Conscious. Stud.* **2012**, *19*, 167–172.
38. Crawford, K. Artificial Intelligence's White Guy Problem. *New York Times*, 25 June 2016.
39. Garling, C. Andrew Ng: Why 'Deep Learning' Is a Mandate for Humans, not just Machines. *Wired*, May 2015.
40. Etzioni, O. No, the Experts Don't Think Superintelligent AI Is a Threat to Humanity. *MIT Technology Review*, 20 September 2016.
41. Dafoe, A.; Russell, S. Yes, We Are Worried about the Existential Risk of Artificial Intelligence. *MIT Technology Review*, 2 November 2016.
42. Baum, S.D. Reconciliation between factions focused on near-term and long-term artificial intelligence. *AI Soc.* **2017**. [CrossRef]
43. Goertzel, B. Superintelligence: Fears, promises and potentials. *J. Evol. Technol.* **2015**, *25*, 55–87.
44. Baum, S.D.; Barrett, A.M.; Yampolskiy, R.V. Modeling and interpreting expert disagreement about artificial superintelligence. *Informatica* **2017**, *41*, 419–428.
45. Bieger, J.; Thórisson, K.R.; Wang, P. Safe baby AGI. In Proceedings of the 8th International Conference on Artificial General Intelligence (AGI), Berlin, Germany, 22–25 July 2015; Bieger, J., Goertzel, B., Potapov, A., Eds.; Springer: Cham, Switzerland, 2015; pp. 46–49.
46. Searle, J.R. What your computer can't know. *The New York Review of Books*, 9 October 2014.
47. Nichols, T. *The Death of Expertise: The Campaign against Established Knowledge and Why It Matters*; Oxford University Press: New York, NY, USA, 2017.
48. De Vrieze, J. 'Science wars' veteran has a new mission. *Science* **2017**, *358*, 159. [CrossRef] [PubMed]
49. Stirling, M. Merchants of Consensus: A Public Battle against Exxon. 2017. Available online: https://ssrn.com/abstract=3029939 (accessed on 18 August 2018).
50. Hampshire, G. Alberta Government Cool on Controversial Climate Change Speaker. *CBC News*, 19 January 2018.
51. Marshall, G. *Don't Even Think About It: Why Our Brains Are Wired to Ignore Climate Change*; Bloomsbury: New York, NY, USA, 2014.
52. Whitehouse, S. *Captured: The Corporate Infiltration of American Democracy*; The New Press: New York, NY, USA, 2017.
53. Kirk, R. Conservation activism is a healthy sign. *Baltimore Sun*, 4 May 1970.

54. Boykoff, M.T.; Boykoff, J.M. Balance as bias: Global warming and the US prestige press. *Glob. Environ. Chang.* **2004**, *14*, 125–136. [CrossRef]
55. Baum, S.D.; Haqq-Misra, J.D.; Karmosky, C. Climate change: Evidence of human causes and arguments for emissions reduction. *Sci. Eng. Ethics* **2012**, *18*, 393–410. [CrossRef] [PubMed]
56. Oreskes, N. The scientific consensus on climate change. *Science* **2004**, *306*, 1686. [CrossRef] [PubMed]
57. CNA Military Advisory Board. *National Security and the Threat of Climate Change*; The CNA Corporation: Alexandria, VA, USA, 2007.
58. Lighthill, J. *Artificial Intelligence: A Paper Symposium*; Science Research Council: Swindon, UK, 1973.
59. Menzies, T. 21st-century AI: Proud, not smug. *IEEE Intell. Syst.* **2003**, *18*, 18–24. [CrossRef]
60. Bentley, P.J. The three laws of artificial intelligence: Dispelling common myths. In *Should We Fear Artificial Intelligence? In-Depth Analysis*; Boucher, P., Ed.; European Parliamentary Research Service, Strategic Foresight Unit: Brussels, Belgium, 2018; pp. 6–12.
61. Häggström, O. A spectacularly uneven AI report. *Häggström Hävdar*, 30 March 2018.
62. Marcus, G. Artificial intelligence is stuck. Here’s how to move it forward. *New York Times*, 29 July 2017.
63. Bengtsson, B. Pinker is dangerous. *Jag är Här*, 22 October 2017.
64. Häggström, O. The AI meeting in Brussels last week. *Häggström Hävdar*, 23 October 2017.
65. Torres, P. A Detailed Critique of One Section of Steven Pinker’s Chapter “Existential Threats” in Enlightenment Now. Project for Future Human Flourishing Technical Report 2, Version 1.2, 2018. Available online: https://docs.wixstatic.com/ugd/d9aaad_8b76c6c86f314d0288161ae8a47a9821.pdf (accessed on 21 August 2018).
66. Clifford, C. Google billionaire Eric Schmidt: Elon Musk is ‘exactly wrong’ about A.I. because he ‘doesn’t understand’. *CNBC*, 29 May 2018.
67. Bogost, I. Why Zuckerberg and Musk are fighting about the robot future. *The Atlantic*, 27 July 2017.
68. Cass, O. The problem with climate catastrophizing. *Foreign Affairs*, 21 March 2017.
69. Cass, O. How to worry about climate change. *National Affairs*, Winter 2017; pp. 115–131.
70. Baum, S.D. The great downside dilemma for risky emerging technologies. *Phys. Scr.* **2014**, *89*, 128004. [CrossRef]
71. Heath, A. Mark Zuckerberg’s plan to create non-voting Facebook shares is going to trial in September. *Business Insider*, 4 May 2017.
72. Ingram, M. At Alphabet, there are only two shareholders who matter. *Fortune*, 7 June 2017.
73. Broockman, D.; Ferenstein, G.F.; Malhotra, N. *The Political Behavior of Wealthy Americans: Evidence from Technology Entrepreneurs*; Stanford Graduate School of Business: Stanford, CA, USA, 2017; Stanford Graduate School of Business Working Paper, No. 3581.
74. Edsall, T.B. Silicon Valley takes a right turn. *New York Times*, 12 January 2017.
75. Mullins, B. Paying professors: Inside Google’s academic influence campaign. *Wall Street Journal*, 15 July 2017.
76. Taplinaug, J. Google’s disturbing influence over think tanks. *New York Times*, 30 August 2017.
77. Tiku, N. New America chair says Google didn’t prompt critic’s ouster. *Wired*, 6 September 2017.
78. Marquis, C.; Toffel, M.W.; Zhou, Y. Scrutiny, norms, and selective disclosure: A global study of greenwashing. *Organ. Sci.* **2016**, *27*, 483–504. [CrossRef]
79. Mack, E. Why Elon Musk spent $10 million to keep artificial intelligence friendly. *Forbes*, 15 January 2015.
80. Pickard, V. Media failures in the age of Trump. *Political Econ. Commun.* **2017**, *4*, 118–122.
81. Lewandowsky, S.; Gignac, G.E.; Vaughan, S. The pivotal role of perceived scientific consensus in acceptance of science. *Nat. Clim. Chang.* **2013**, *3*, 399–404. [CrossRef]
82. Cook, J.; Lewandowsky, S. The Debunking Handbook. St. Lucia, Australia: University of Queensland, 2011. Available online: https://skepticalscience.com/Debunking-Handbook-now-freely-available-download.html (accessed on 18 August 2018).
83. Chan, M.P.; Jones, C.R.; Hall Jamieson, K.; Albarracín, D. Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychol. Sci.* **2017**, *28*, 1531–1546. [CrossRef] [PubMed]
84. Slovic, P. The perception gap: Radiation and risk. *Bull. At. Sci.* **2012**, *68*, 67–75. [CrossRef]

information

Essay

The Singularity Isn't Simple! (However We Look at It) A Random Walk between Science Fiction and Science Fact

Vic Grout

Department of Computing, Wrexham Glyndŵr University, Wrexham LLI1 2AW, UK.; v.grout@glyndwr.ac.uk; Tel.: +44-1978-293-203

Received: 11 April 2018; Accepted: 18 April 2018; Published: 19 April 2018

Abstract: It seems to be accepted that *intelligence—artificial* or otherwise—and 'the singularity' are inseparable concepts: 'The singularity' will apparently arise from AI reaching a, supposedly particular, but actually poorly-defined, level of sophistication; and an empowered combination of hardware and software will take it from there (and take over from us). However, such wisdom and debate are simplistic in a number of ways: firstly, this is a poor definition of the singularity; secondly, it muddles various notions of intelligence; thirdly, competing arguments are rarely based on shared axioms, so are frequently pointless; fourthly, our models for trying to discuss these concepts at all are often inconsistent; and finally, our attempts at describing any 'post-singularity' world are almost always limited by anthropomorphism. In all of these respects, professional 'futurists' often appear as confused as storytellers who, through freer licence, may conceivably have the clearer view: perhaps then, that becomes a reasonable place to start. There is no attempt in this paper to propose, or evaluate, any research hypothesis; rather simply to challenge conventions. Using examples from science fiction to illustrate various assumptions behind the AI/singularity debate, this essay seeks to encourage discussion on a number of possible futures based on different underlying metaphysical philosophies. Although properly grounded in science, it eventually looks beyond the technology for answers and, ultimately, beyond the Earth itself.

Keywords: futurism and futurology; hard science fiction; artificial intelligence; models of consciousness; intelligent machines; machine replication; machine evolution and optimization; technological singularity

1. Introduction: Problems with 'Futurology'

This paper will irritate some from the outset by treating 'serious' academic researchers, professional 'futurists' and science fiction writers as largely inhabiting the same techno-creative space. Perhaps, for these purposes, our notion of science fiction might be better narrowed to 'hard' sci-fi (i.e., that with a supposedly realistic edge) loosely set somewhere in humanity's (rather than anyone else's) future; but, that aside, the alignment is intentional, and no apology is offered for it.

Moreover, it is justified to a considerable extent: purveyors of hard future-based sci-fi and professional futurism have much in common. They both have their work (their credibility) judged in the here-and-now by known theory and, retrospectively, by empirical evidence. In fact, there may be a greater gulf between (for example) the academic and commercial futurist than between either of them and the sci-fi writer. Whereas the professional futurists may have their objectives set by technological or economic imperatives, the storyteller is free to use a scientific premise of their choosing as a blank canvas for *any* wider social, ethical, moral, political, legal, environmental or demographic discussion. This 360° view is becoming increasingly essential: asking technologists their view on the future of technology makes sense; asking their opinions regarding its wider impact may not.

1.1. Futurology 'Success'

Therefore, for our purposes, we define an alternative role, that of *'futurologist'* to be any of these, whatever their technical or creative background or motivation may be. A futurologist's predictive success (or *accuracy*) may be loosely assessed by their performance across three broad categories: *positives*, *false positives* and *negatives*, defined as follows:

- *Positives*: predictions that have (to a greater or lesser extent) come to pass within any suggested time frame [*the futurologist predicted it and it happened*];
- *False positives*: predictions that have failed to transpire or have only done so in limited form or well beyond a suggested time frame [*the futurologist predicted it but it didn't happen*];
- *Negatives*: events or developments either completely unforeseen within the time frame or implied considerably out of context [*the futurologist didn't see it coming*].

Obviously, these terms are vague and subject to overlap, but they are for discussion only: there will be no attempt to quantify them here as metrics. (However, see [1] for an informal attempt at doing just this!) Related to these is the concept of justification: the presence (or otherwise) of a coherent argument to support predictions. Justification may be described not merely by its presence, partial (perhaps unconvincing) presence or absence but also by its form or nature; purely scientific, for example, or based on wider social, economic, legal, etc. grounds. The ideal would be a set of predictions with high accuracy (perhaps loosely the ratio of positives to false positives and negatives) together with strong justification. The real world, however is never that simple. Although unjustified predictions might be considered merely guesses, some guesses are lucky: another reason why the outcomes of factual and fictional writing may not be so diverse. Finally, assessment of prediction accuracy and justification rarely happen together. Predictions made *now* can be considered for their justification but not their accuracy: that can only be known later. Predictions made in the past can have their positives and negatives scrutinised in detail but their justification has been made, by the passage of time and the comfort of certainty, less germane.

This flexible assessment of accuracy (positives, false positives and negatives) and justification can be applied to any attempt at futurology, whatever its intent: to inform, entertain or both. We start then with one of the best known examples of all.

Star Trek [2,3] made its TV debut in 1966. Although five decades have now passed, we are still over two centuries from its generally assumed setting in time. This makes most aspects of assessment of its futurology troublesome but, as a snapshot exercise/example in the here-and-now, we can try:

- *Positives*: (as of 2018) voice interface computers, tricorders (now in the form of mobile phones), Bluetooth headsets, in-vision/hands-free displays, portable memory units (now disks, USB sticks, etc.), GPS, tractor beams and cloaking devices (in limited form), tablet computers, automatic doors, large-screen/touch displays, universal translators, teleconferencing, transhumanist bodily enhancement (bionic eyes for the blind, etc.), biometric health and identity data, diagnostic hospital beds [4];
- *False Positives*: (to date) replicators (expected to move slowly to positive in future?) warp drives and matter-antimatter power, transporters, holodecks (slowly moving to positive?), the moneyless society, human colonization of other planets [5];
- *Negatives*: (already) The Internet? The Internet of Things? Body Area Networks (BANs)? Personal Area Networks (PANs)? Universal connectivity and coverage? [6] (*but all disputed* [7]);
- *Justification*: pseudo-scientific-technological in part but with social, ethical, political, environmental and demographic elements?

A similar exercise can be (indeed, informally has been) attempted with other well-known sci-fi favourites such as Back to the Future [8] and Star Wars [9] or, with the added complexity of insincerity, Red Dwarf [10].

1.2. Futurology 'Failure'

In Star Trek's case, the interesting, and contentious, category is the negatives. Did it really fail to predict modern, and still evolving, networking technology? Is there no Internet in Star Trek? There are certainly those who would defend it against such claims [11] but such arguments generally take the form of noting secondary technology that could *imply* a pervasive global (or universal) communications network: there is no first-hand reference point anywhere. The ship's computer, for example, clearly has access to a vast knowledge base but this always appears to be held locally. Communication is almost always point-to-point (or a combination of such connections) and any notion of distributed processing or resources is missing. Moreover, in many scenes, there is a plot-centred (clearly intentional) sense of isolation experienced by its characters, which is incompatible with today's understanding and acceptance of a ubiquitous Internet: on that basis alone, it is unlikely that its writers intended to suggest one.

This point regarding (overlooking) 'negatives' may be more powerfully made by considering another sci-fi classic. Michael Moorcock's *Dancers at the End of Time* trilogy [12] was first published in full in 1976 and has been described as *"one of the great postwar English fantasies"* [13]. It describes an Earth millions of years into the future in which virtually anything is possible. Through personal 'power rings', its cast of decadent eccentrics have access to almost limitless capability from technology established in the distant past by long-lost generations of scientists. As the hidden technology consumes vast, remote stellar regions for its energy, the inhabitants of this new world can create life and cheat death. It is an excellent example of the use of a simple sci-fi premise as a starting point for a wider social and moral discussion. Setting the scene takes just a few pages but then it gets interesting: *how do people behave* in a world where *anything* is possible?

Yet, there is no Internet; or even anything much by way of mobile technology. If 'The Dancers' want to observe what might be happening on the other side of the planet, they use their power rings to create a plane (or a flying train) and go there at high speed: fun, perhaps, but hardly efficient! A mere two decades before the Internet became a reality for most people in the developed world, Moorcock's vision of an ultimately advanced technological civilization did not include such a concept.

Of course, many writers from E.M. Forster (in 1928) [14], through Will Jenkins (1946) [15], to Isaac Asimov (1957) [16], even Mark Twain (much earlier in 1898) [17], have described technology with Internet-like characteristics so there can be no concluding that imagining any particular scientific innovation is necessarily impossible. However each of these, whilst pointing the way in this particular respect to varying degrees, are significantly adrift in other areas. In some there is an archaic reliance on existing equipment; in others, a network of sorts but no computers. Douglas Adams's *Hitchhiker's Guide to the Galaxy* (1978) [18] is a universal knowledge base without a network. There is no piece of futurology from more than two or three decades ago which portrays the current world particularly accurately in even most, let alone all, respects. Not only does technological advancement have too many individual threads, the interaction between them is hugely complex.

This is not a failing of fiction writers only. To give just one example, in 1977, Ken Olsen, founder and CEO of DEC [19], made an oft-misapplied statement, *"there is no reason for any individual to have a computer in his home"*. A favourite of introductory college computing modules, supposedly highlighting the difficulty in keeping pace with rapid change in computing technology, it appears foolish at a time when personal computers were already under development, including in his own laboratories. The quote is out of context, of course, and applies to Olsen's scepticism regarding fully-automated assistive home technology systems (climate control, security, cooking food, etc.) [20]. However, as precisely these technologies gain traction, there may be little doubt that, if he stands unfairly accused of being wrong in one respect, time will inevitably prove him so in another. This form of 'it won't happen' prediction (but it then does) can be considered simply as an extension of the *negatives* category for futurology success or, if necessary, as an additional false negatives set; but this would digress this discussion unnecessarily at this point.

Therefore, the headline conclusion of this introduction is, of course, the somewhat obvious 'futurology is difficult' [21]. However, there are three essential, albeit intersecting, components of this observation, which we can take forward as the discussion shifts towards the eponymous technological singularity:

- *Advances in various technologies do not happen independently or in isolation*: Prophesying the progress made by artificial intelligence in ten years' time may be challenging; similarly, the various states-of-the art in robotics, personal communications, the Internet of Things and big data analytics are difficult to predict with much confidence. However, combine *all* these and the problem becomes far worse. Visioning an 'always-on', fully-connected, fully-automated world driven by integrated data and machine intelligence superior to ours is next to impossible;
- *A dominant technological development can have a particularly unforeseen influence on many others*: Moorcock's technological paradise may look fairly silly with no Internet but the 'big thing that we're not yet seeing' is an ever-present hazard to futurologists. As we approach (the possibility of) the singularity, what else (currently concealed) may arise to undermine the best intentioned of forecasts?
- *Technological development has wider influences and repercussions than merely the technology*: Technology does not emerge or exist in a vacuum: its development is often driven by social or economic need and it then has to take its place in the human world. *We* produce it but it then changes *us.* Wider questions of ethics, morality, politics and law may be restricting or accelerating influences but they most certainly cannot be dismissed as irrelevant.

These elements have to remain in permanent focus as we move on to the main business of this paper.

2. Problems with Axioms, Definitions and Models

Having irritated some traditional academics at the start of the previous section, we begin this one in a similar vein: by quoting Wikipedia [22].

> *"The technological singularity (also, simply, the singularity) [23] is the hypothesis that the invention of artificial superintelligence will abruptly trigger runaway technological growth, resulting in unfathomable changes to human civilization [24]. According to this hypothesis, an upgradable intelligent agent (such as a computer running software-based artificial general intelligence) would enter a "runaway reaction" of self-improvement cycles, with each new and more intelligent generation appearing more and more rapidly, causing an intelligence explosion and resulting in a powerful superintelligence that would, qualitatively, far surpass all human intelligence. Stanislaw Ulam reports a discussion with John von Neumann "centered on the accelerating progress of technology and changes in the mode of human life, which gives the appearance of approaching some essential singularity in the history of the race beyond which human affairs, as we know them, could not continue" [25]. Subsequent authors have echoed this viewpoint [22,26]. I. J. Good's "intelligence explosion" model predicts that a future superintelligence will trigger a singularity [27]. Emeritus professor of computer science at San Diego State University and science fiction author Vernor Vinge said in his 1993 essay The Coming Technological Singularity that this would signal the end of the human era, as the new superintelligence would continue to upgrade itself and would advance technologically at an incomprehensible rate [27].*
>
> *At the 2012 Singularity Summit, Stuart Armstrong did a study of artificial general intelligence (AGI) predictions by experts and found a wide range of predicted dates, with a median value of 2040 [28].*
>
> *Many notable personalities, including Stephen Hawking and Elon Musk, consider the uncontrolled rise of artificial intelligence as a matter of alarm and concern for humanity's future [29,30]. The consequences of the singularity and its potential benefit or harm to the human race have been hotly debated by various intellectual circles."*

Informal as much of this is, it provides a useful platform for discussion and, although there may be more credible research bases, no single source will be the consensus of such a large body of opinion: in this sense, it can be argued to represent at least the common view of the Technological Singularity (TS). Once again, we see sci-fi suggestions of the TS (Vinge) pre-dating academic discussions (Kurzweil [31]) and consideration of the wider social, economic, political and ethical impact also follows close behind: these various dimensions cannot be considered in isolation; or, at least, should not be.

2.1. What Defines the 'Singularity'?

However, before embarking on a discussion of what the TS actually is, if it might happen, how it might take place and what the implications could be, a more fundamental question merits at least passing consideration: is the TS really a 'singularity' at all?

This is not an entirely simple question to answer. Different academic subjects, from various fields in mathematics, through the natural sciences, to emerging technologies have their own concept of a 'singularity' with most dictionary definitions imprecisely straddling several of them. Sci-fi [27] continues to contribute in its own fashion. However, a common theme is the notion of the rules, formulae, behaviour or understanding of any system breaking down or not being applicable at the point of singularity. ('We understand how the system works everywhere except here') A black hole is a singularity in space-time: the conventional laws of physics fail as gravitational forces become infinite (albeit infinitely slowly as we approach it). The function $y = 1/x$ has a singularity at $x = 0$: it has no (conventional) value. On this basis, *is* the point in our future, loosely understood to be that at which machines start to evolve by themselves, really a singularity? To put it another way, does it suggest the sort of discontinuity most definitions of a singularity imply?

It may not. Whilst there may be little argument with there being a period of great uncertainty following on from the TS, including huge questions regarding what the machines might do or how long (or just how) humans, or the planet, might survive [31], it is not clear that the TS marks an impassable break in our timeline. It is absolutely unknown what happens to material that enters a black hole; it may appear in some other form elsewhere in the universe but that is entirely speculative: there is no other side, of which we know. The $y = 1/x$ curve can be followed smoothly towards $x = 0$ from either (positive or negative) direction but there is no passing continuously through to the other. However, at the risk of facetiousness, if we go to bed one night and the TS occurs while we sleep, we will still most likely wake up the following morning. The world may indeed have become a very different place but it will probably continue, for a while at least. It is possible that we sometimes confuse 'uncertainty' with 'discontinuity' so perhaps 'singularity' is not an entirely appropriate term?

Returning, to less abstract matters, we attempt to consider what the TS might actually be. Immediately, this draws us into difficulty as we begin to note that most definitions, including the Wikipedia consensus, conflate different concepts:

- *Process-based*: (forward-looking) We have a model of what technological developments are required. When each emerges and combines with the others, the TS should happen: 'The TS will occur *when* ... '; [This can also be considered as a *'white box'* definition: we have an idea of how the TS might come about];
- *Results-based*: (backward-looking) We have a certain outcome or expectation of the TS. The TS has arrived when technology passes some *test*: 'The TS will have occurred when ... '; [Also a *'black box'* definition: we have a notion of functionality but not (necessarily) operation];
- *Heuristic*: ('rule-of-thumb', related models) Certain developments in technology, some themselves defined better than others, will continue along (or alongside) the road towards the TS and may contribute, directly or indirectly, to it: 'The TS will occur as ... '; [Perhaps a *'grey box'* definition?].

Figure 1 attempts to distil these overlapping principles in relation to the TS. Towards the left of the diagram, more is understood (with clearer, concrete definitions) and more likely to achieve consensus among futurologists. Towards the right, concepts become more abstract, with looser definitions and less agreement. Time moves downwards.

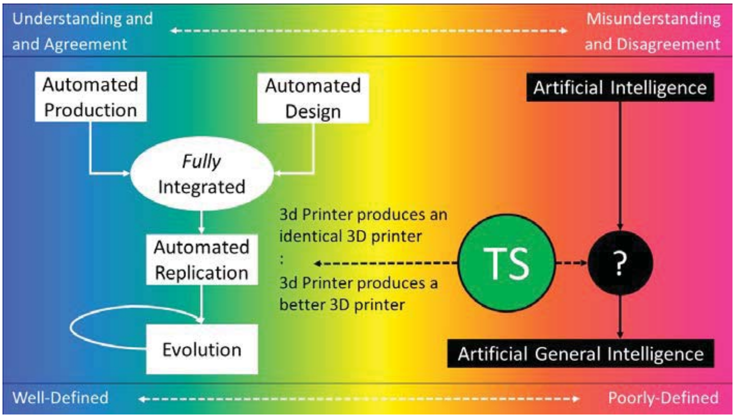


Figure 1. Combined (but simplified) models of the technological singularity. TS, Technological Singularity.

2.2. How Might the 'Singularity' Happen?

Any definition of the TS by reference to AI, itself not well defined, or worse, to *AGI* (*Artificial General Intelligence*: loosely the evolutionary end-point of AI to something 'human-like' [32]) is, at best, a heuristic one. Not only does this not provide a deterministic characterization of the TS, it is not *axiomatically* dependent upon it. Suppose there was an *agreed* formulaic measurement for AI (or AGI), which there is not, what *figure* would we need to reach for the TS to occur? We have no idea and it is unlikely that the question is a functional one. In addition, *is* it unambiguously clear that this is the *only* way it could happen? Not if we base our definitions on precise *expectations* rather than *assumptions*.

A process-based model (white box), built on what we already have, is at least easier to understand. *Evolution* is the result of iterated replication within a 'shaping' or 'guiding' environment. In theory, at least, (automatic) replication is the automated combining of *design* and *production*. We already see examples of both automated production and automated design at work. Production lines, including robotic hardware, have been with us for decades [33]: in this respect, accelerated manufacturing began long ago. Similarly, software has become an increasingly common feature of engineering processes over the years [34] with today's algorithms being very sophisticated indeed [35]. Many problems in electronic component design, placement, configuration and layout [36,37], for example, would be entirely impossible now without computers taking a role in optimizing their successors. In principle, it is merely a matter of connecting all of this seamlessly together!

Fully integrating automated design and automated production into automated replication is far from trivial, of course, but it does give a better theoretical definition of the TS than vague notions of AI or AGI. It is, however, also awash with practical problems:

1 [Design] Currently, the use of design software and algorithms is piecemeal in most industries. Although some of the high-intensity computational work is naturally undertaken by automated procedures, there is almost always human intervention and mediation between these sub-processes. It varies considerably from application to application but nowhere yet is design automation close to comprehensive in terms of a complete 'product'.

2 [Production] A somewhat more mundane objection: but even an entirely self-managed production line has to be fed with materials; or components perhaps if production is multi-stage, but raw materials have to start the process. If this part of the TS framework is to be fully automated

and *sustainable* then either the supply of materials has to be similarly seamless or the production hardware must somehow be able to source its own. Neither option is in sight presently.

3 [Replication] Fully integrating automated design with production is decidedly non-trivial. Although there are numerous examples of production lines employing various types of feedback, to correct and improve performance, this is far short of the concept of taking a *final* product design produced by software and immediately implementing it to hardware. Often, machinery has to be reset to implement each new design or human intervention is necessary in some other form. Whilst 1 and 2 require considerable technological advance, this may ask for something more akin to a human 'leap of faith'.

4 [Evolution] Finally, if 1, 2 and 3 were to be checked off at some point in the future (*not* theoretically impossible), the transition from static replication to a chain of improvements across subsequent iterations requires a further ingredient: *the guiding environment, or purpose, required for evolution across generations*. Where would this come from? Suppose it requires some extra program code or an adjustment to the production hardware to produce the necessary improvement at each generation, how would (even could) such a decision be taken and what would be the motivation (of the hardware and software) for taking it? Couched in terms of an optimization problem [38], what would the objective function be and where would the responsibility lie to improve it?

Although consideration of 4 does indeed appear to lead us back to some notion of intelligence, which is why the heuristic (grey box) definition of the TS associated with AI or AGI is not outright wrong, it is simply unhelpful in any practical sense. Exactly *what* level of *what* measurement of intelligence would be necessary for a self-replicating system to effect improvement across generations? It remains unclear. Having broken a process-based TS down into its constituents, however, it does suggest an alternative definition.

Because a results-based (black box) definition of the TS could be something much easier to come to terms with. For example, we might understand it to be the point at which a 3D printer autonomously creates an identical version of itself. Then, presumably, with some extra lines of code, it creates a marginally better version. This superior offspring has perhaps slightly better hardware features (precision, control, power, features or capabilities) and improved software (possibly viewed in terms of its 'operating system'). As the child becomes a parent, further hardware and software enhancements follow and the evolutionary process is underway. Even here we could dispute the *exact* point of the TS but it is clearly framed somewhere in the journey from the initial parent, through its extra code, to the first child. 'Intelligence' has not been mentioned.

The practical objections above still remain, of course. We can perhaps ignore 1 [Design] by reminding ourselves that this is a black box definition: we can simply claim to recognize when the outcomes have satisfied our requirements. Similarly 2 [Production] might be overlooked but it does suggest that the 3D printer might have to have an independent/autonomous supply network of (not just present but all possible future) materials required: *very* challenging, arguably *non-deterministic*; the alternative, though, might be that the device has to be *mobile*, allowing it to source or forage for its needs! Finally, although we can use the black box defence against being able to describe 3 [Replication] in detail, the difficult question of 4 [Evolution] remains: *Why would the printer want to do this?* Therefore, this might bring us back to intelligence after all; or does it?

Because, if we are forced to deal with intelligence as a driver towards the TS (even if only to attempt to answer the question of whether it is even required), we encounter a similar problem in what may be a recurring pattern: we have very little idea (understanding or agreement) on what *'intelligence'* is ether, which takes us to the next section.

3. Further Problems with 'Thinking', 'Intelligence', 'Consciousness', General Metaphysics and 'The Singularity'

The irregular notion of an intelligent machine is found in literature long before academic non-fiction. Although it may stretch definitions of 'sci-fi', *Talos* [39] appears in stories dating back to the third century BC. A giant bronze automaton, patrolling the shores of Europa (Crete), he appears to have many of the attributes we would expect of an intelligent machine. Narratives vary somewhat but most describe him as of essentially non-biological origin. Many other examples follow over the centuries with Mary Shelley's monster [40] being one of the best known: Dr. Frankenstein's creation, however, is the result of (albeit corrupted) natural science. These different versions of ('natural' or 'unnatural') 'life' or 'intelligence' continue apace into the 21st century, of course [41–44], but, by the time Alan Turing [45] asked, in 1950, whether such things were possible in practice, sci-fi had already given copious judgement in theory.

How important is this difference between 'natural' and 'unnatural' life? Is it the same as the distinction between 'artificial' and 'real' intelligence? Who decides what is natural and unnatural anyway? (The concepts of 'self-awareness' and 'consciousness' are discussed later.) As Turing pointed out tangentially in his response to 'The Theological Objection [45], this may not be a question we have the right, or ability, to make abstract judgements on. (To paraphrase his entirely atheist argument: 'even if there were a God, who are *we* to tell Him what he can and cannot do?') However, even in the title of this seminal work ('Computing Machinery and Intelligence'), followed by his opening sentence (*'I propose to consider the question, "Can machines think?"'*), he has apparently already decided that 'thinking' and 'intelligence' are the same thing. Although he immediately follows this with an admission that precise definitions are problematic, whether this equivalence should be regarded is axiomatic is debatable.

3.1. 'Thinking' Machines

However, something that Turing [45] forecasts with unquestionable accuracy is that, *'I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted'*. It cannot be denied that today we do just this (and have done for some time): at the very least, a computer is 'thinking' while we await its response. It might be argued that the use of the word in this context describes merely the delay, or even that it has some associated irony attached, but we still do it, even if we do not give any thought to what it means. Leaving aside his almost universally misunderstood 'Imitation Game' ('Turing Test': see later) discussion, in this the simplest of predictions, Turing was right.

(His next sentence is then often overlooked: *'I believe further that no useful purpose is served by concealing these beliefs. The popular view that scientists proceed inexorably from well-established fact to well-established fact, never being influenced by any improved conjecture, is quite mistaken. Provided it is made clear which are proved facts and which are conjectures, no harm can result. Conjectures are of great importance since they suggest useful lines of research.'* This very much supports the guiding principles of this paper: so long as we can recognize the difference between science-fact and anything short of this, conjecture, and even fiction, have value in promoting discussion. Perhaps the only real difference between the sci-fi writer and the professional futurist is that the latter expect us to believe them?)

3.2. 'Intelligent' Machines

Therefore, if 'thinking' is a word used too loosely to have much value, what of the perhaps stronger 'intelligence'? Once again, the dictionary is of little help, confusing several clearly different concepts. 'Intelligence', depending on the context, implies a capacity for: *learning*, *logic*, *problem-solving*, *reasoning*, *creativity*, *emotion*, *consciousness* or *self-awareness*. That these are hardly comparable is more than abstract semantics: does 'intelligent' simply suggest 'good at something' (number-crunching, for example) or is it meant to imply 'human-like'? By some of these definitions, a pocket-calculator is intelligent, a chess program more so. However, by others, a fully-automated humanoid robot, faster, stronger and superior to its creators in every way would not be if it lacked consciousness and/or self-awareness [46]. What is meant (in Star Trek, for example) by frequent references to 'intelligent life'? Is this tautological or are there forms of life lacking intelligence? (After all, we often say this of people we lack respect for.) Are the algorithms that forecast the weather or run our economies intelligent? (They do things with a level of precision that we cannot.) If intelligence can be anything from simple learning to full self-awareness, at which point does AI become AGI, or perhaps 'artificial' intelligence become 'real' intelligence. There sometimes seem to be as many opinions as there are academics/researchers. And, once again, we return to what increasingly looks to be a poorly framed question: *how much 'intelligence' is necessary for the TS?*

Perhaps tellingly, Turing himself [45] made no attempt to define intelligence, either at human or machine level. In fact, with respect to the latter, accepting his conflation of intelligence with thought, he effectively dismisses the question: he precedes his *'end of the century'* prediction above by, *'The original question, "Can machines think?" I believe to be too meaningless to deserve discussion.'* Instead, he proposes his, widely quoted but almost as widely misunderstood, *'Imitation Game'*, eventually to become known as the 'Turing Test'. A human 'interrogator' is in separate, remote communication with both a machine and another human and, through posing questions and considering their responses, has to decide which is which. It probably makes little difference whether either is allowed to lie. (Clearly then something akin to 'intellectually human-like' is thus being suggested as intelligence here.) In principle, if the interrogator simply cannot decide (or makes the wrong decision as often as the right one) then the machine has 'won' the game.

Whatever the thoughts of sci-fi over the decades [47,48] might be, this form of absolute victory was not considered by Turing in 1950. Instead he made a simple (and quite possibly off-the-cuff) prediction that *'I believe that in about fifty years' time it will be possible, to programme computers, . . . , to make them play the imitation game so well that an average interrogator will not have more than 70 per cent chance of making the right identification after five minutes of questioning.'* Today's reports [49] of software 'passing the Turing Test' because at least four people out of ten chose wrongly in a one-off experiment are a poor reflection on this original discussion. We should also consider that the sophistication of the interrogators and their questioning may well increase over time: both pocket calculators and chess machines seemed unthinkable shortly before they successfully appeared but are now mainstream.

However, the key point regarding this formulation of an 'intelligence test' is that it is only loosely based on knowledge. No human knows nothing but some know more than others. Similarly, programming a computer with knowledge bases of various sizes is trivial. The issue for the interrogator is to attempt to discriminate on the basis of how that knowledge is manipulated and presented. 'Q. What did Henry VIII die from'; 'A. Syphilis', requires a simple look-up but 'Q. Name a British monarch who died from a sexually transmitted disease', is more complex since neither the king nor the illness is mentioned in the original question/answer. As a slightly different illustration, asking for interpretations of the signs, 'Safety hats must be worn', and 'Dogs must be carried', is challenging: although syntactically similar, their regulatory expectations are very different.

3.3. 'Adaptable' Machines

This, in turn, leads us to the concept of adaptability, often mooted in relation to intelligence, by both human-centric and AI researchers. Some even take this further to the point of using it to define 'genius' [50]. The argument is that *what* an entity *knows* is less important than how it *uses* it, particularly in dealing with situations where that knowledge has to be applied in a different context to that in which it was learned. This relates well to the preceding paragraph and perhaps has a ring of truth in assessing the intelligence of humans or even other animals; but what of AI? What does adaptability look like in a machine?

In fact, this is very difficult and rests squarely on what we think might happen as machines approach, then evolve through, the TS. If our current models of software running on hardware (as separate entities) survive, then adaptability may simply be the modification of database queries to analyse existing knowledge in new ways or an extension to existing algorithms to perform other tasks. Thus, in a strictly evolutionary sense, it may, for example, be sufficient for one generation to autonomously acquire better software in order to effect hardware improvements in the next. This is essentially a question of extra code being written, which is becoming simple enough, even automatically [51], if we (again) put aside the question of what the machine's motivation would be to do this.

All these approaches to machine 'improvement', however, remain essentially under human control. If, in contrast, adaptability in AI or machine terms implies fundamental changes to how that machine actually operates (perhaps, but not necessarily, within the same generation), or we expect the required 'motivation for improvement' to be somehow embedded in this evolutionary process, then this is entirely different, and may well have to appear independently. The same conceptual difficulty arises if we look to machines in which hardware and software are merged into an inseparable operational unit, as they appear to be in our human brains [52]. If the future has machines such as this: machines that have somehow passed over a critical evolutionary boundary, beyond anything we have either built or programmed into them, then something very significant indeed will have happened. What might cause this? To an extent, our accepted framework within which machines/computers operate may have been replaced by something conceptually beyond our control (possibly our understanding), one of the deeper concerns regarding the TS. However, where would this come from? Well, there may be numerous possibilities, so this is by no means an irrefutable logical progression we follow now but, having postponed the discussion to this point, it seems appropriate to deal with *'consciousness'* or *'self-awareness'*.

3.4. 'Conscious' Machines

Not all writers and academics, from neuroscientists to philosophers [53], consider these terms entirely synonymous but the distinction is outside the scope of this paper. However, whether consciousness necessarily implies intelligence, particularly how that might play out in practice, is certainly contentious and returning to sci-fi illustrates this point well. If, only temporarily, we follow a (largely unproven) line of reasoning that consciousness is related to 'brain' size, that is if we create sufficiently complex hardware or software then sentience emerges naturally, then, rather than look to stand-alone machines or even arrays of supercomputers, our first sight of such a phenomenon would presumably come from the single largest (and most complex) thing that humanity has ever created: the *Internet*. Whilst, in this context, Terminator's 'Skynet' [54] is well known, the considerable variants on this theme are best illustrated by two lesser-known novels.

Robert J. Sawyer's *WWW* Trilogy [41–43] describes the emergence of *'Webmind'*, an AI apparently made from discarded network packets endlessly circling the Internet topology. Leaving aside the question of whether the 'time-to-live' field [55] typically to be found in such packets might be expected to clear such debris, the story suggests an interesting framework for emergent consciousness based purely on software. Sawyer loosely suggests that a large enough number of such packets, interacting in the nature of cellular automata [42] would eventually achieve sentience. For any real evidence we have to the contrary, perhaps they could; there is nothing fundamentally objectionable in Sawyer's pseudo-science.

However, *Webmind* exhibits an equally serious problem we often observe in trying to discuss AI, probably AGI, in fact, or any post-TS autonomous computer or machine: the tendency to be unrealistically anthropocentric, or worse. In *WWW*, not only is *Webmind* intelligent by any reasonable definition, it is also almost limitlessly powerful. Once it has mastered its abilities, it is quickly able to 'do good' as it sees it, partially guided by a (young) human companion. It begins by eliminating electronic spam, then finds a cure for cancer and finally overthrows the Chinese government. It would remain to be seen, of course, quite how 'good' any all-powerful AI would really regard each of these objectives. At best, the first is merely a high-level feature of human/society, the second would presumably improve human lives but the overall balance of nature is unknown (even if the AI would really care much about *that*) and the last is the opinion of a financially-comfortable, liberally-minded western author: not *all humans* would necessarily share this view, let alone an emergent intelligence! Frankly, whatever A(G)I may perceive its raison d'être to be, it is unlikely to be this!

In marked contrast, in the novel, *Conscious*, Vic Grout's *'It'* [44], far from being software running on the Internet, is simply the Internet *itself*. The global combination of the Internet and its supporting power networks have reached such a level of scale and connection density, they automatically acquire an internal control imperative and begin behaving as a (nascent form of) independent 'brain'. Hardware and software appear to be working inseparably: although its signals can be observed, they emanate from nowhere other than the hardware devices and connections themselves. Humans have supplied all this hardware, of course, and although, initially, *It* works primitively with human software protocols, these are slowly adapted to its own (generally unfathomable) purposes. *It* begins its life as the physical (wired) network infrastructure before subsuming wireless and cellular communications as well. Eventually, *It* acquires total mastery of the *'Internet of Everything'* (the *'Internet of Things'* projected a small number of years into the future) and causes huge, albeit random, damage.

Crucially, though, Grout quite deliberately offers no insight into the workings of *Its* 'mind'. Although *It* is almost immeasurably powerful (its peripherals are effectively everything on Earth: human transport, communications, climate control, life-support, weaponry, etc.), it remains throughout no more than embryonic in its sophistication, even its intelligence. If it *can* be considered brain-like, it is a brain in the very earliest stages of development. *It* is undoubtedly *conscious* in the sense that it has acquired an independent control imperative, and, as the story unfolds, there is clear evidence of learning: at least, becoming familiar with its own structure and (considerable) capabilities, but in no sense could this be confused with being 'clever'. There is no human attempt to communicate with *It* and not the slightest suggestion that this could be possible. Although *It* eventually wipes out a considerable fraction of life on Earth, it is unlikely that it understands any of this in any conventional sense. *It* is as much simply a 'powered nervous system' as it is a brain. *Conscious* [44] is only a story, of course, but, as a model, if actual consciousness could indeed be achieved without real intelligence (perhaps it requires merely a simple numerical threshold, neural complexity, for example, to be reached) then it would be difficult to make any confident assertions regarding AI or AGI.

3.5. Models of 'Consciousness'

Grout's framework for a conscious Internet is essentially a variant of the philosophical principle of panpsychism [56]. (Although 'true' panpsychism suggests some level of universal consciousness in both living and non-living things, Grout's model suggests a need for external 'power' or 'fuel' and an essential 'tipping point', neural complexity here, at which consciousness either happens or becomes observable.) The focus of any panpsychic model is essentially based on hardware. However, as with Sawyer's model of a large software-based AI, for all we really know of any of this, it is as credible as any other.

As with 'intelligence', there are almost as many models of 'consciousness' [53] as there are researchers in the field. However, loosely grouping the main themes together [57], allows us to make a final, hopefully interesting, point. To this end, the following could perhaps be considered the main 'types' of explanations of where consciousness come from:

1. Consciousness is simply the result of neural complexity. Build something with a big enough 'brain' and it will acquire consciousness. There is possibly some sort of critical neural mass and/or degree of complexity/connectivity for this to happen.
2. Similar to 1 but the 'brain' needs energy. It needs power (food, fuel, electricity, etc.) to make it work (Grout's 'model' [44]).
3. Similar to 2 but with some symbiosis. A physical substrate is needed to carry signals of a particular type. The relationship between the substrate and signals (hardware and software) takes a particular critical form (maybe the two are indistinguishable) and we have yet to fully grasp this.
4. Similar to 3 but there is a biological requirement. Consciousness is the preserve of carbon life forms, perhaps. How and/or why we do not yet understand, and it is conceivable that we may not be able to.

5. A particular form of 4. Consciousness is somehow separate from the underlying hardware but still cannot exist without it (Sawyer's 'model' [41–43]?).
6. Extending 5. Consciousness is completely separate from the body and could exist independently. Dependent on more fundamental beliefs, we might call it a 'soul'.
7. Taking 6 to the limit? Consciousness, the soul, comes from God.

Firstly, before dismissing any of these models out-of-hand, we should perhaps reflect that, if pressed, a non-negligible fraction of the world's population would adopt each of them (which is relevant if we consider *attitudes* towards AI). Secondly, each is credible in two respects: (1) they are logically and separately non-contradictory and (2) the human brain, which we assume to deliver consciousness, could (for all we know) be modelled by any of them. In considering each of these loose explanations, we need to rise above pre-conceptions based on belief (or lack of). In fact, it may be more instructive to look at *patterns*. With this in mind, these different broad models of consciousness can be viewed as being in sequence in two senses:

- In a simple sense, they (Models 1–7) could be considered as ranging from the 'ultra-scientific' to the 'ultra-spiritual'.
- Increasingly sophisticated machinery will progressively (in turn) achieve some of these. 1 and 2 may be trivial, 3 a distinct possibility, and 4 is difficult to assert. After that, it would appear to get more challenging. In a sense, we might disprove each model by producing a machine that satisfied the requirements at each stage, yet failed to 'wake up'.

However, the simplicity of the first 'progression' is spoilt by adding 'pure' panpsychism and the second by Turing himself.

Firstly, if we add the pure panpsychism model as a new item zero in the list:

Everything in the universe has consciousness to some extent, even the apparently inanimate: humans are simply poor when it comes to recognizing this, then this fits nicely before (as a simpler version of) 1 whilst, simultaneously, the panpsychic notion of a 'universal consciousness' looks very much like an extension of 9. (It closely matches some established religions.) The list becomes cyclic and notions of 'scientific' vs. 'spiritual' are irreparably muddled.

Secondly, whatever an individual's initial position on all of this, Turing's original 1950 paper [45] could perhaps challenge it . . .

3.6. A Word of Caution . . . for Everyone

In suggesting (loosely) that some form of A(G)I might be possible, Turing anticipated 'objections' in a variety of forms. The first, he represented thus:

> *"The Theological Objection: Thinking is a function of man's immortal soul. God has given an immortal soul to every man and woman, but not to any other animal or to machines. Hence no animal or machine can think."*

He then dealt with this robustly as follows:

> *"I am unable to accept any part of this, but will attempt to reply in theological terms. I should find the argument more convincing if animals were classed with men, for there is a greater difference, to my mind, between the typical animate and the inanimate than there is between man and the other animals. . . . It appears to me that the argument quoted above implies a serious restriction of the omnipotence of the Almighty. It is admitted that there are certain things that He cannot do such as making one equal to two, but should we not believe that He has freedom to confer a soul on an elephant if He sees fit? We might expect that He would only exercise this power in conjunction with a mutation which provided the elephant with an appropriately improved brain to minister to the needs of this sort. An argument of exactly similar form may be made for the case of machines. It may seem different*

> *because it is more difficult to 'swallow'. But this really only means that we think it would be less likely that He would consider the circumstances suitable for conferring a soul. . . . In attempting to construct such machines we should not be irreverently usurping His power of creating souls, any more than we are in the procreation of children: rather we are, in either case, instruments of His will providing mansions for the souls that He creates."*

Turing is effectively saying (to 'believers'), 'Who are you to tell God how His universe works?'

This is a significant observation in a wider context, however. Turing's caution, although clearly derived from a position of unequivocal atheism, can perhaps be considered as a warning to any of us who have already decided what is possible or impossible in regard to A(G)I, then expect emerging scientific (or other) awareness to support us. That is not the way anything works; neither science nor God are extensions of ourselves: they will not rally to support our intellectual cause because we ask them to. It makes no difference what domain we work in; whether it is (any field of) science or God, or both, that we believe in, it is *that* that will decide what can and cannot be done, not *us*.

In fact, we could at least tentatively propose the notion that, ultimately, this is something we *cannot* understand. There is nothing irrational about this: most scientific disciplines from pure mathematics [58], through computer science [59] to applied physics [60] have known results akin to 'incompleteness theorems'. We are entirely reconciled to the acceptance that, starting from known science, there are things that cannot be reached by human means: propositions that cannot be proved, problems that cannot be solved, measurements that cannot be taken, etc. However, with a median predicted date little more than 20 years away [22], can the reality of the TS really be one of these undecidables? How? Well, perhaps we may simply never get to find out . . .

4. The Wider View: Ethics, Economics, Politics, the Fermi Paradox and Some Conclusions

The *'Fermi Paradox'* (*FP*) [61] can be paraphrased as asking, on the cascaded assumptions that there are a quasi-infinite number of planets in the universe, with a non-negligible, and therefore still extremely large, number of them supporting life, why have none of them succeeded in producing civilizations sufficiently technologically advanced to make contact with us? Many answers are proposed [62] from the perhaps extremes of *'there isn't any: we're alone in the universe'* to suggestions of *'there* has *been contact: they're already here, we just haven't noticed'*, etc.

It could be argued or implied, however, that the considerable majority of possible solutions to the FP suggest that *the TS has not happened on any other planet in the universe*. Whatever, the technological limitations of its 'natural' inhabitants may have been, whatever may or may not have happened to them (survival or otherwise) following their particular TS, whatever physical constraints and limitations they were unable to overcome, their new race(s) of intelligent machines, evolving at approaching an infinite rate, would surely be visible by now. As they do not appear to be, we are forced into a limited number of possible conclusions. Perhaps the machines have no desire to make contact (not even with the other machines they presumably know must exist elsewhere) but, again, this (machine motivation) is simply unknown territory for us. Perhaps these TSs are genuinely impossible for all the reasons discussed here and elsewhere [63]. Or perhaps their respective natural civilizations never get quite that far.

4.1. Will We Get to See the TS?

This, indeed, is an FP explanation gaining traction as our political times look more uncertain and our technological ability to destroy ourselves increases on all fronts. It has been suggested [64] that it may be a natural (and unavoidable) fate of advanced civilizations that they self-destruct before they are capable of long-distance space travel or communication. Perhaps this inevitable doom also prevents any TS? We should probably not discount this theory lightly. Aside from the obvious weaponry and the environmental damage being done, we have other, on the surface, more mundane, factors at work. Social media's ability, for example, to threaten privacy, spread hatred, strew false information, cause division, etc., or the spectre of mass unemployment through automation, could all destabilize society to crisis point.

Technology has the ability to deliver a utopian future for humanity (with or without the TS occurring) but it almost certainly will not, essentially because it will always be put to work within a framework in which the primary driving force is profit [65]. A future in which intelligent machines do all the work for us could be good for everyone; but for it to be, fundamental political-economic structures would have to change, and there is no evidence that they are going to. (Whether 'not working' for a human is a good thing has nothing to do with the technology that made their work unnecessary but the economic framework everything operates under. If nothing changes, 'not working' will be socially unpleasant, as it is now: technology cannot change that, but there could soon be more out of work than in.) Under the current political-economic arrangement, nothing really happens for the general good, whatever public-facing veneer may be applied to the profit-centric system. However, such inequality and division may eventually (possibly even quickly) lead to conflict? The framework itself, however, appears to be very stable so it may be that it preserves itself to the bitter end. All in all, the outlook is a bleak one. Perhaps the practical existence of the TS is genuinely unknowable because it lies just beyond any civilization's self-destruct point, so it can never be reached?

4.2. Can We Really Understand the TS?

Coming back to more Earthly and scientific matters, we should also, before ending, note something essential regarding evolution itself: it is imperfect [66]. The evolutionary process is, in effect, a sophisticated local-search optimization algorithm (variants are applied to current solutions in the hope of finding better ones in terms of a defined objective). The objective (for biological species, at least) is survival and the variable parameters are these species' characteristics. The evolutionary algorithm can never guarantee 'perfection' since it can only look (for each successive generation) at relatively small changes in its variables (mutations). As Richard Dawkins puts it [67] in attempting to deal with the question of *'why don't animals have wheels?'*

> *"Sudden, precipitous change is an option for engineers, but in wild nature the summit of Mount Improbable can be reached only if a gradual ramp upwards from a given starting point can be found. The wheel may be one of those cases where the engineering solution can be seen in plain view, yet be unattainable in evolution because it lies the other side of a deep valley, cutting unbridgeably across the massif of Mount Improbable."*

Therefore, advanced machines, seeking evolutionary improvements, may not achieve perfection. Does this mean they will remain constrained (and 'content') with imperfection or will/can they find a better optimization process? There are certainly optimization algorithms superior to local-search (often somewhat cyclically employing techniques we might loosely refer to as AI) and this/these new races(s) of superior machines might realistically have both the processing power and logic to find them. In the sense we generally apply the term, machines may not *evolve* at all: they may do something *better*. There may or may not be the same objective. Millions of years of evolution could, perhaps, be bypassed by a simple calculation? However, whatever happens, the chances of us being able to understand it are minimal.

4.3. Therefore, Can the TS Really Happen?

Returning finally to the main discussion, if we persist in insisting that the TS can only arise from the evolution of AI to AGI, or worse that machines have to achieve self-awareness for this to happen then we are beset by difficulties: we really have little idea what this really means so a judgement on whether it can happen is next to impossible. There are credible arguments [68] that machines can never achieve the same form of intelligence as humans and that Turing's models [41] are simplistic in this respect. If both the AGI requirement for the TS and its (AGI's) impossibility to achieve are simultaneously correct then, trivially, the TS cannot happen. However, if the TS can result simply from the conjunction of increasingly complex, but fully-understood, technological processes, then perhaps it can: and a model of what this might look like in practical terms is, not only possible but, independent of definitions of intelligence. Finally, other predictions of various crises in humanity's relationship with technology and its wider impact [65] within the framework of existing social, economic and political practice could yet render the debate academic.

However, in many ways, the real take home message is that many of us are not on the same page here. We use terms and discuss concepts freely with no standardization as to what they mean; we make assumptions in our self-contained logic based on axioms we do not share. Whether across different academic disciplines, wider fields of interest, or simply as individuals, we have to face up to some uncomfortable facts in a very immediate sense: that many of us are not discussing the same questions. On that basis it is hardly surprising that we seem to be coming to different conclusions. If, as appears to be the case, a majority of (say) neuroscientists think the TS cannot happen and a majority of computer scientists think it can then, assuming an equivalent distribution of intelligence and abilities over those disciplines, clearly they are not visioning the same event. *We need to talk.*

One final thought though, an admission really, as this paper has been written, probably in common with many others in a similar vein, by a middle-aged man having worked through a few decades of technological development: sometimes with appetite, but at other times with horror ... could this possibly be just a 'generational thing'? Just as Bertrand Russell [69] noted that philosophers themselves are products of, and therefore guided by, their position and period, futurologists are unlikely to be (at best) any better. This may yet all make more sense to generations to come. There is no better way to conclude a paper, considering uncertain futures and looking to extend thought boundaries through sci-fi, than to leave the final word to Douglas Adams [70].

> *"I've come up with a set of rules that describe our reactions to technologies:*
>
> 1. *Anything that is in the world when you're born is normal and ordinary and is just a natural part of the way the world works.*
> 2. *Anything that's invented between when you're fifteen and thirty-five is new and exciting and revolutionary and you can probably get a career in it.*
> 3. *Anything invented after you're thirty-five is against the natural order of things."*

Conflicts of Interest: The author declares no conflict of interest.

References

1. Can Futurology 'Success' Be 'Measured'? Available online: https://vicgrout.net/2017/10/11/can-futurology-success-be-measured/ (accessed on 12 October 2017).
2. Goodman, D.A. *Star Trek Federation*; Titan Books: London, UK, 2013, ISBN 978-1781169155.
3. Siegel, E. *Treknology: The Science of Star Trek from Tricorders to Warp Drive*; Voyageur Press: Minneapolis, MN, USA, 2017, ISBN 978-0760352632.
4. Here Are All the Technologies Star Trek Accurately Predicted. Available online: https://qz.com/766831/star-trek-real-life-technology/ (accessed on 20 September 2017).
5. Things Star Trek Predicted Wrong. Available online: https://scifi.stackexchange.com/questions/100545/things-star-trek-predicted-wrong (accessed on 20 September 2017).
6. Mattern, F.; Floerkemeier, C. From the Internet of Computers to the Internet of Things. *Inf. Spektrum* **2010**, *33*, 107–121. [CrossRef]
7. How Come Star Trek: The Next Generation did not Predict the Internet? Is It Because It Will Become Obsolete in the 24th Century? Available online: https://www.quora.com/How-come-Star-Trek-The-Next-Generation-did-not-predict-the-internet-Is-it-because-it-will-become-obsolete-in-the-24th-century (accessed on 20 September 2017).
8. From Hoverboards to Self-Tying Shoes: Predictions that Back to the Future II Got Right. Available online: http://www.telegraph.co.uk/technology/news/11699199/From-hoverboards-to-self-tying-shoes-6-predictions-that-Back-to-the-Future-II-got-right.html (accessed on 20 September 2017).

9. The Top 5 Technologies That 'Star Wars' Failed to Predict, According to Engineers. Available online: https://www.inc.com/chris-matyszczyk/the-top-5-technologies-that-star-wars-failed-to-predict-according-to-engineers.html (accessed on 21 September 2017).
10. How Red Dwarf Predicted Wearable Tech. Available online: https://www.wareable.com/features/rob-grant-interview-how-red-dwarf-predicted-wearable-tech (accessed on 21 September 2017).
11. Star Trek: The Start of the IoT? Available online: https://www.ibm.com/blogs/internet-of-things/star-trek/ (accessed on 21 September 2017).
12. Moorcock, M. *Dancers at the End of Time*; Gollancz: London, UK, 2003, ISBN 978-0575074767.
13. When Hari Kunzru Met Michael Moorcock. Available online: https://www.theguardian.com/books/2011/feb/04/michael-moorcock-hari-kunzru (accessed on 19 September 2017).
14. Forster, E.M. *The Machine Stops*; Penguin Classics: London, UK, 2011, ISBN 978-0141195988.
15. Leinster, M. *A Logic Named Joe*; Baen Books: Wake Forest, CA, USA, 2005, ISBN 978-0743499101.
16. Asimov, I. *The Naked Sun*; Panther/Granada: London, UK, 1973, ISBN 978-0586010167.
17. Twain, M. *From the 'London Times' of 1904*; Century: London, UK, 1898.
18. Adams, D. *The Hitch Hiker's Guide to the Galaxy: A Trilogy in Five Parts*; Heinemann: London, UK, 1995, ISBN 978-0434003488.
19. Rifkin, G.; Harrar, G. *The Ultimate Entrepreneur: The Story of Ken Olsen and Digital Equipment Corporation*; Prima: Roseville, CA, USA, 1990, ISBN 978-1559580229.
20. Did Digital Founder Ken Olsen Say There Was 'No Reason for Any Individual to Have a Computer in His Home'? Available online: http://www.snopes.com/quotes/kenolsen.asp (accessed on 22 September 2017).
21. The Problem with 'Futurology'. Available online: https://vicgrout.net/2013/09/20/the-problem-with-futurology/ (accessed on 26 October 2017).
22. Technological Singularity. Available online: https://en.wikipedia.org/wiki/Technological_singularity (accessed on 16 January 2018).
23. Cadwalladr, C. Are the robots about to rise? Google's new director of engineering thinks so .... *The Guardian (UK)*, 22 February 2014.
24. Eden, A.H.; Moor, J.H. *Singularity Hypothesis: A Scientific and Philosophical Assessment*; Springer: New York, NY, USA, 2013, ISBN 978-3642325601.
25. Ulam, S. Tribute to John von Neumann. *Bull. Am. Math. Soc.* **1958**, *64*, 5.
26. Chalmers, D. The Singularity: A philosophical analysis. *J. Conscious. Stud.* **2010**, *17*, 7–65.
27. Vinge, V. The Coming Technological Singularity: How to survive the post-human era. In *Vision-21 Symposium: Interdisciplinary Science and Engineering in the Era of Cyberspace*; NASA Lewis Research Center and Ohio Aerospace Institute: Washington, DC, USA, 1993.
28. Armstrong, S. How We're Predicting AI. In *Singularity Conference*; Springer: San Francisco, CA, USA, 2012.
29. Sparkes, M. Top scientists call for caution over artificial intelligence. *The Times (UK)*, 24 April 2015.
30. Hawking: AI could end human race. *BBC News*, 2 December 2014.
31. Kurzweil, R. *The Singularity is Near: When Humans Transcend Biology*; Duckworth: London, UK, 2006, ISBN 978-0715635612.
32. Everitt, T.; Goertzel, B.; Potapov, A. *Artificial General Intelligence*; Springer: New York, NY, USA, 2017, ISBN 978-3319637020.
33. Storr, A.; McWaters, J.F. *Off-Line Programming of Industrial Robots: IFIP Working Conference Proceedings*; Elsevier Science: Amsterdam, The Netherlands, 1987, ISBN 978-0444701374.
34. Aouad, G. *Computer Aided Design for Architecture, Engineering and Construction*; Routledge: Abingdon, UK, 2011, ISBN 978-0415495073.
35. Liu, B.; Grout, V.; Nikolaeva, A. Efficient Global Optimization of Actuator Based on a Surrogate Model. *IEEE Trans. Ind. Electron.* **2018**, *65*, 5712–5721. [CrossRef]
36. Liu, B.; Irvine, A.; Akinsolu, M.; Arabi, O.; Grout, V.; Ali, N. GUI Design Exploration Software for Microwave Antennas. *J. Comput. Des. Eng.* **2017**, *4*, 274–281. [CrossRef]
37. Wu, M.; Karkar, A.; Liu, B.; Yakovlev, A.; Gielen, G.; Grout, V. Network on Chip Optimization Based on Surrogate Model Assisted Evolutionary Algorithms. In Proceedings of the IEEE Congress on Evolutionary Computation (CEC), Beijing, China, 6–8 July 2014; pp. 3266–3271.
38. Sioshansi, R.; Conejo, A.J. *Optimization in Engineering: Models and Algorithms*; Springer: New York, NY, USA, 2017, ISBN 978-3319567679.

39. Rhodius, A. *The Argonautica*; Reprinted; CreateSpace: Seattle, WA, USA, 2014, ISBN 978-1502885616.
40. Shelley, M. *Frankenstein: Or, the Modern Prometheus*; Reprinted; Wordsworth: Ware, UK, 1992, ISBN 978-1853260230.
41. Sawyer, R.J. *WWW: Wake*; The WWW Trilogy Part 1; Gollancz: London, UK, 2010, ISBN 978-0575094086.
42. Sawyer, R.J. *WWW: Watch*; The WWW Trilogy Part 2; Gollancz: London, UK, 2011, ISBN 978-0575095052.
43. Sawyer, R.J. *WWW: Wonder*; The WWW Trilogy Part 3; Gollancz: London, UK, 2012, ISBN 978-0575095090.
44. Grout, V. *Conscious*; Clear Futures Publishing: Wrexham, UK, 2017, ISBN 978-1520590127.
45. Turing, A.M. Computing Machinery and Intelligence. *Mind* **1950**, *49*, 433–460. [CrossRef]
46. How Singular Is the Singularity? Available online: https://vicgrout.net/2015/02/01/how-singular-is-the-singularity/ (accessed on 16 February 2018).
47. Dick, P.K. *Do Androids Dream of Electric Sheep*; Reprinted; Gollancz: London, UK, 2007, ISBN 978-0575079939.
48. Poelti, M. *A.I. Insurrection: The General's War*; CreateSpace: Seattle, WA, USA, 2018, ISBN 978-1981490585.
49. Computer AI Passes Turing Test in 'World First'. Available online: http://www.bbc.co.uk/news/technology-27762088 (accessed on 19 February 2018).
50. Adaptability: The True Mark of Genius. Available online: https://www.huffingtonpost.com/tomas-laurinavicius/adaptability-the-true-mar_b_11543680.html (accessed on 16 February 2018).
51. Our Computers Are Learning How to Code Themselves. Available online: https://futurism.com/4-our-computers-are-learning-how-to-code-themselves/ (accessed on 19 February 2018).
52. Why Your Brain Isn't a Computer. Available online: https://www.forbes.com/sites/alexknapp/2012/05/04/why-your-brain-isnt-a-computer/#3e238fdc13e1 (accessed on 19 February 2018).
53. Schneider, S.; Velmans, M. *The Blackwell Companion to Consciousness*, 2nd ed.; Wiley-Blackwell: Hoboken, NJ, USA, 2017, ISBN 978-0470674079.
54. August 29th: Skynet Becomes Self-Aware. Available online: http://www.neatorama.com/neatogeek/2013/08/29/August-29th-Skynet-Becomes-Self-aware/ (accessed on 23 March 2018).
55. Time-to-Live (TTL). Available online: http://searchnetworking.techtarget.com/definition/time-to-live (accessed on 23 March 2018).
56. Seager, W. *The Routledge Handbook of Panpsychism*; Routledge Handbooks in Philosophy; Routledge: London, UK, 2018, ISBN 978-1138817135.
57. "The Theological Objection". Available online: https://vicgrout.net/2016/03/06/the-theological-objection/ (accessed on 27 March 2018).
58. Gödel, K. Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme, I. *Mon. Math. Phys.* **1931**, *38*, 173–198. (In German)
59. Turing, A.M. On computable numbers, with an application to the Entscheidungsproblem. *Proc. Lond. Math. Soc.* **1937**, *42*, 230–265. (In German) [CrossRef]
60. Heisenberg, W. Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik. *Z. Phys.* **1927**, *43*, 172–198. (In German)
61. Popoff, A. *The Fermi Paradox*; CreateSpace: Seattle, WA, USA, 2015, ISBN 978-1514392768.
62. Webb, S. *If the Universe is Teeming with Aliens . . . Where is Everybody? Fifty Solutions to the Fermi Paradox and the Problem of Extraterrestrial Life*; Springer: New York, NY, USA, 2010, ISBN 978-1441930293.
63. Braga, A.; Logan, R.K. The Emperor of Strong AI Has No Clothes: Limits to Artificial Intelligence. *Information* **2017**, *8*, 156. [CrossRef]
64. Do Intelligent Civilizations Across the Galaxies Self Destruct? For Better and Worse, We're the Test Case. Available online: http://www.manyworlds.space/index.php/2017/02/01/do-intelligent-civilizations-across-the-galaxies-self-destruct-for-better-and-worse-were-the-test-case/ (accessed on 28 March 2018).
65. Technocapitalism. Available online: https://vicgrout.net/2016/09/15/technocapitalism/ (accessed on 16 February 2018).
66. The Algorithm of Evolution. Available online: https://vicgrout.net/2014/02/03/the-algorithm-of-evolution/ (accessed on 2 April 2018).
67. Why don't Animals Have Wheels? Available online: http://sith.ipb.ac.rs/arhiva/BIBLIOteka/270ScienceBooks/Richard%20Dawkins%20Collection/Dawkins%20Articles/Why%20don%3Ft%20animals%20have%20wheels.pdf (accessed on 2 April 2018).

68. Logan, R.K. Can Computers Become Conscious, an Essential Condition for the Singularity? *Information* **2017**, *8*, 161. [CrossRef]

69. Russell, B. *History of Western Philosophy*; reprinted; Routledge Classics: London, UK, 2004, ISBN 978-1447226260.

70. Adams, D. *The Salmon of Doubt: Hitchhiking the Galaxy One Last Time*; reprinted; Pan: London, UK, 2012, ISBN 978-0415325059.

information

Article

Pareidolic and Uncomplex Technological Singularity

Viorel Guliciuc *

Department of Human, Social and Political Sciences, Faculty of History and Geography, "Ştefan cel Mare" University, 13 University Street, 720229 Suceava, Romania

Received: 25 October 2018; Accepted: 3 December 2018; Published: 6 December 2018

Abstract: "Technological Singularity" (TS), "Accelerated Change" (AC), and Artificial General Intelligence (AGI) are frequent future/foresight studies' themes. Rejecting the reductionist perspective on the evolution of science and technology, and based on *patternicity* ("the tendency to find patterns in meaningless noise"), a discussion about the perverse power of *apophenia* ("the tendency to perceive a connection or meaningful pattern between unrelated or random things (such as objects or ideas)") and *pereidolia* ("the tendency to perceive a specific, often meaningful image in a random or ambiguous visual pattern") in those studies is the starting point for two claims: *the "accelerated change" is a future-related* apophenia *case*, whereas *AGI (and TS) are future-related* pareidolia *cases*. A short presentation of research-focused social networks working to solve complex problems reveals the superiority of human networked minds over the hardware-software systems and suggests the opportunity for a network-based study of TS (and AGI) from a complexity perspective. It could compensate for the weaknesses of approaches deployed from a linear and predictable perspective, in order to try to redesign our intelligent artifacts.

Keywords: Technological Singularity; Accelerated Change; Artificial (General) Intelligence; apophenia; pareidolia; complexity; research focused social network; networked minds; complexity break; complexity fallacy

"Any fact becomes important when it's connected to another"

(Umberto Eco, *Foucault's Pendulum*)

1. A Pretext

The popular understanding of the Future(s) and, especially, of AGI and/or of the TS seems to be related *to patternicity* [1], *apophenia* [2] and *pareidolia* [3].

We care about the future because, as living beings, we are "programmed" for the conservation of our lives despite threats, challenges, and changes.

When the future has become an essential part of their lives *and* their language(s), pre-human beings have become human beings.

Change, the possibility of change, and the power-of-realization/of-becoming-real of the possibility/virtuality-of-change are essential parts of our relationship with reality, with what- is.

Because nothing can be without the power-to-be, the power to erupt from what-is-virtual into what-is-real [4] (pp. 91–92), this power-of-being/power-to-be is fueling both our future(s) and our studies of the future(s).

2. The Fascination with the Future: Foresight Studies Require an Appropriate Methodology

Confronted with the challenge of correctly understanding, representing, and managing the future(s) of humankind and, especially, future discoveries in science and breakthroughs in technology, we have, primarily, to correctly deal with our perplexities and expectations related to them, and to find the most probable behavior of such complex systems as the human brain/mind, human society, science and/or technology for our best future(s).

Let us remember that, when "the simple yet infinitely complex question of where is technology taking us?" [5] was asked, our times were described as the "Age of Surprise."

It is a "concept originally described by the U.S. Air Force Center for Strategy and Technology at The Air University" [5], a part of the Blue Horizons project [6], writes Reuven Cohen.

In fact, the Age of Surprise is a situation in which "the exponential advancement of technology" has just reached "a critical point where not even governments can project the direction humanity is headed" [5].

What is interesting for this paper's theme is the forecast made by some researchers of "an eventual Singularity where the lines between humans and machines are blurred" [5].

There is increasing awareness related to technological changes and breakthroughs.

Under this metaphor, but also considering the necessity of finding the best solutions in order to not engage humankind in catastrophic, global, and existential risks [7] and, especially and more specifically, in *existential technological risks*, it is essential, from a scientific approach, *to not accept that it is natural to find what we expect to find*, *nor to project our expected finds as sound scientific results*. Yet, it is also important *to not reduce the complexity of the analyzed systems—TS or AGI—to the linearity of a predictable use of data*, when dealing with the future(s).

In fact, *to date, there is no such* scientific field *as future studies* [8]/*foresight studies*, even though there are such claims [9].

Because future-related reasoning is probabilistic, a hard science of the future is problematic, as we cannot decide with any accuracy about the truth or falsehood of our judgments on future actions, or on the existence of future beings and future artifacts. As one of the reviewers of this paper rightly observed, in science we make predictions that can be "evaluated according to available experience" and this is a sign that "nevertheless, the science is possible." Indeed, a probabilistic truth engages a re-evaluation of the classic two-dimensional models of reasoning in favor of a discussion related to an unified model of reasoning [10].

The sense of the above revised statement is related to some of the observations made by Samuel Arbesman in his book *The Half-life of Facts: Why Everything We Know Has an Expiration Date* [11]. Even though I could not access the contents of the book, I have understood, from the review by Roger M. Stein, that *in science, acquired knowledge has a pace of overturn* [12], and, from the paper of James A. Evans on "Future Science," that *innovation has been decreasing, instead of accelerating, in the last several decades* [13].

This is why it is difficult and problematic to claim the status of a "hard" science for foresight studies.

Perhaps the explanation for such a problematic status could be related not only to the complexity of the future-related models of the evolution of science and technology (as proven in "The Age of Surprise" report), but, also, to Aristotle's legacy [14] regarding the problem of future contingents [15] and/or to Charles Sanders Peirce's observation that, in order to reach, through inquiry, an objective scientific understanding we have to eliminate such human factors as expectations and bias ([16] pp. 111–112).

A major issue in the study of future(s) is reaching the best *probably* correct understanding. For such an accomplishment, *we have to accept both the complexity of the world and the complexity of the human being(s)*.

This is why, first of all, we need to reject any reductionist perspective.

There is a powerful reductionist tendency in scientific studies, as "the world is still dominated by a mechanistic, cause and effect mindset with origins in the Industrial Revolution and the Newtonian scientific philosophy" [17].

Beyond the radical claim in the sentence quoted above, the existence of such reductionist tendencies is of importance for TS researches, as such tendencies are detectable in various perspectives on AC, AGI, and TS as well. This is why, as one of the reviewers concluded, this paper is focused on "criticism of the TS/AC/AGI claims based on reductionism and extrapolation."

In the following pages, I will briefly consider some examples of the perverse power of our expectations related to the future(s) of technology and, especially related to TS, I will engage in a short discussion based on the observation that there are few studies on TS deployed from a complexity perspective.

3. The Apophenic Face(s) of Our Technology- and Science-Related Expectations: The "Law of Accelerating Returns" (LAR)

Contemplating the timeline of humankind, when change is perceived as affecting both societies and individuals, debates and controversies related to the future flourish until the reaching of a new *equilibrium* and a new *status quo* regarding the perception(s) and understanding(s) of the possible future(s)—not necessarily predictable and predicted in those discussions.

a. *Is the Change in Science and Technology Accelerating?*

One of the most appealing ideas in the debates of our times is so-called "accelerating change" (AC)/the "law of accelerating returns" (LAR).

When considering the idea of AC, one should go back to the idea of "progress itself" [18].

This is not about AC in technology only—e.g., as claimed by Ray Kurzweil [19]. It is about AC in science, too—e.g., as claimed by John M. Smart, the Foresight U, and the FERN teams [18].

For some researchers, AC is an ontological feature, almost a law of nature as it "is one of the most future-important, pervasive, and puzzling features of our universe," everywhere observable, including "the twenty-thousand year history of human civilization" [18].

Meanwhile, for other researchers it emerges universally [20].

In general terms, AC is just a perceived change of rhythm in the regularity of the advances of technology and science through the ages.

The challenge it brings with it, observes Richard Paul, is "to trace the general implications of what are identified as the two central characteristics of the future: accelerating change and intensifying complexity" [21]. The major concerns are related to the pace of AC and to the perception of an increasing complexity: "how are we to understand how this change and complexity will play itself out? How are we to prepare for it?" [21].

The fascination with AC is an expression of our interest in the study of the future, in so-called "foresight studies" [22].

In fact, it is such a rapidly growing topic that, using the simplest Google search, on 4 October 2018, I found not only about 16,800,000 results for "foresight" about 11,700,000 results for "foresight studies," and about 14,900,000 results for "foresight study," but also 20,400 results for "foresight science," too.

At the same time, *the perceived AC seems to be more likely just a subjective projection of our assumptions/presuppositions*, even when discovering the gap(s) between our linear and predictable expectations about the pace(s) of changes in technology and science and our own minds' power to process the complexity of the information related to the new developments in science and new breakthroughs in technology.

Some critics of AC—Theodore Modis [23]; Jonathan Huebner [24]; Andrey Korotayev, Artemy Malkov, and Daria Khaltourina [25]; James A. Evans [13]; Julia Lane [26] and bloggers such as Richard Jones [27] and David Moschella [28], among others—argued that "the rate of technological innovation has not only ceased to rise, but is actually now declining" [29] and/or argued that there are other possible projections of the LAR (Law of Accelerating Returns) besides the one proposed by Kurzweil.

To date, there is neither general acceptance of AC's existence, nor a final rejection of it.

Under these circumstances, a statement such as "one can criticize 'proofs' of AC for a subjective selection of technologies, but no one can claim that within the selective set of technologies there is no AC" (made by one of the reviewers of this paper) is most likely a subjective one.

My main concerns are: How is it/could it be/should it be established that a "selective set of technologies" should be generally accepted? and Is it permitted to extrapolate those particular findings to a universal set of technologies or even to the status of a universal phenomenon?

In the meantime, it is well known that there are several models of Singularity and of TS—such as those presented in the taxonomies of John Smart [30] and Anders Sandberg [31], for example.

In John Smart's classification, all three types of Singularity (computational, developmental, and technological) "assume the proposition of the universe-as-a-computing-system" [30]. In fact, it is an "assumption, also known as the "infopomorphic" paradigm," that "proposes that information processing is both a process and purpose intrinsic to all physical structures," when "the interrelationships between certain information processing structures can occasionally undergo irreversible, accelerating discontinuities" [30].

I think this is why Amnon H. Eden, James H. Moor, Johnny H. Søraker, and Eric Steinhart chose as the title of the book they co-edited on TS: *Singularity Hypotheses: A Scientific and Philosophical Assessment* [32].

Because, for now, the "infomorphic" paradigm/assumption is still debated and because this is the ground/ultimate source of the TS hypothesis, the results of the debates on TS cannot be, consequently, boldly and clearly related to truth or falsehood.

As TS is deeply related to AC—it is not possible without it!—a good method for an appropriate TS study is to remember the claims of one of the most well-known defenders of TS, Kurzweil.

John Smart is considered among the "prominent explorers and advocates of the technological Singularity idea," along with brilliant researchers such as John Von Neumann, I.J. Good, Hans Moravec, Vernor Vinge, Danny Hillis, Eliezer Yudkowsky, Damien Broderick, Ben Goertzel and "a small number of other futurists, and most eloquently to date, Ray Kurzweil" [30].

For that "eloquence," I choose to refer here, especially, to Kurzweil.

(In the meantime, the "infopomorphic" assumption remains active for TS's defenses).

One of Kurzweil's main ideas is that change is exponential and not "intuitively linear"—as is the Kurzweil's case, he thinks, even with "sophisticated commentators" who "extrapolate the current pace of change over the next ten years or one hundred years to determine their expectations." Meanwhile, "a serious assessment of the history of technology" will reveal that "technological change is exponential" for "a wide variety of technologies, ranging from electronic to biological . . . the acceleration of progress and growth applies to each of them" [33].

However, despite Kurzweil's belief/beliefs—a "natural" result of his own expectations—it was observed that our world is changing, but not as fast as we would be tempted to think, based on our own perceptions: "the world is changing. But change is not accelerating." So, the very idea of an exponential growth of change is disputable [34].

b. Has the Road toward TS a Single Shape/Route?

A second main idea of Kurzweil's is related to the *inevitability of TS*. It is based on a subjective extension, following Moravec, of the so-called "Moore's law" of exponential growth in computing power, having the shape of an asymptote, as a graphic representation of AC, toward a human-like AGI's existence and accelerated growth.

Even in popular science it was observed that the so-called LAR has "altered public perception of Moore's law," because, contrary to the common belief promoted by Kurzweil and Moravec, Moore is *only* making predictions related to the performance(s) of "semiconductor circuits" and not "predictions regarding all forms of technology" [29].

In fact, there were numerous and various debates on the very existence of Kurzweil's extension of Moore's law [35,36]. In the references we cite just four of them.

Most likely, it is not possible to claim exponential growth in the form of a vertical-like asymptote where the curve approaches +∞ [37] (Figure 1).

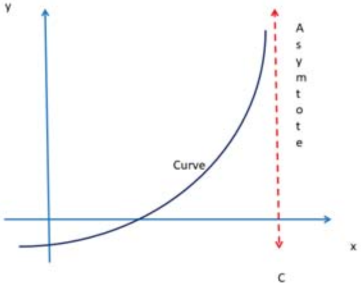


Figure 1. Vertical asymptote as in the following source: [37].

(One of this paper's reviewers noted "the exponential growth doesn't have a vertical asymptote." Indeed, but *this asymptotic-like shape seems to be used by Kurzweil* in order to represent the growth of computing power as in Figure 2 toward a vertical endless growth. More likely the cause is quite simple: it is better at fitting *his expectations* and shows the patterns *he* was looking for. As Korotayev observes, Kurzweil uses *three graphs* to illustrate the countdown to Singularity [38] (page 75). Yet he argues that Kurzweil did not know about the mathematical Singularity, nor was he aware of the differences between exponential and hyperbolic growth, etc.

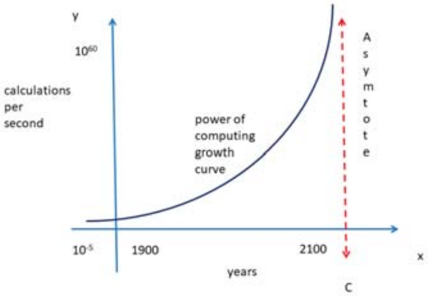


Figure 2. The exponential scale of technological growth (as in [19]).

But, instead, in the form of a horizontal one (Figure 3),

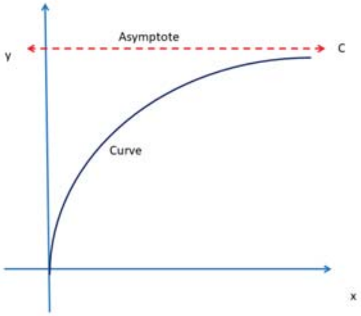


Figure 3. Horizontal asymptote as in the source: [39].

Or even, more naturally, in the form of successive horizontal asymptotes, but vertically arranged (Figure 4),

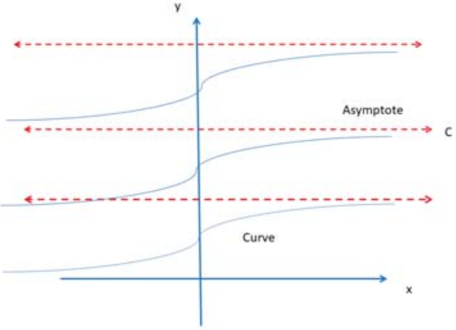


Figure 4. Horizontal asymptotes [39].

I think the Figure 4 fits more naturally the history of breakthroughs in science and technology, as it evokes a succession of jumps in the evolution towards TS.

Indeed, two asymptotes will constitute a hyperbola under certain conditions.

One of the reviewers observed, "The asymptote in TS appears, because each next type of cybernetic systems develops with a higher exponential growth rate, and the time between metasystem transactions to novel cybernetic systems becomes smaller."

However, Kurzveil's TS is related to the growth of computing power, which is related to his extrapolation of Moore's Law. Under these circumstances, let us remember Tom Simonite's paper from the *MIT Technology Review*, "Moore's Law Is Dead. Now What?", where he quotes Horst Simon, deputy director of the Lawrence Berkeley National Laboratory, saying "the world's most powerful calculating machines appear to be already feeling the effects of Moore's Law's end times. The world's top supercomputers aren't getting better at the rate they used to" [40].

The reviewer continues: "consequently, it doesn't really matter if the exponential growth of individual technologies continues infinitely or becomes S-shaped curve with a horizontal asymptote if new technologies outperform older technologies with higher growth rate."

I have some difficulties in understanding the extrapolation made above.

A first one is related to the observation made by Andrey Korotayev, who states: "let us stress again that the mathematical analysis demonstrates rather rigorously that the development acceleration pattern within Kurzweil's series is NOT exponential (as is claimed by Kurzweil), but hyperexponential, or, to be more exact, hyperbolic" [38] (p. 84).

This is why some researchers, studying the macro-regularities in the evolution of humankind, came to different conclusions about the possible evolution of AC. A second perplexity is related to the observation that it really matters if the exponential growth of individual technologies is a S-shaped curve [23].

Andrey Korotayev's conclusions, in his re-analysis of 21st-century Singularity, are: "the existence of sufficiently rigorous global macroevolutionary regularities (describing the evolution of complexity on our planet for a few billions of years)" are "surprisingly accurately described by extremely simple mathematical functions." Moreover, he thinks, "there is no reason "to expect an unprecedented (many orders of magnitude) acceleration of the rates of technological development" near/in the region of the so-called "Singularity point." Instead, "there are more grounds for interpreting this point as an indication of an inflection point, after which the pace of global evolution will begin to slow down systematically in the long term" [38].

This is why I used the representation within Figure 4 (above). The main idea is not to correctly represent the succession of the exponential growth, but to highlight the discontinuities in the evolution of technology. I agree with Richard Jones, who observes: "the key mistake here is to think that 'Technology' is a single thing, that by itself can have a rate of change, whether that's fast or slow." Indeed, "there are many technologies, and at any given time some will be advancing fast, some will be in a state of stasis, and some may even be regressing." Moreover, "it's very common for technologies to have a period of rapid development, with a roughly constant fractional rate of improvement, until physical or economic constraints cause progress to level off" [27].

So, *the mathematical representation of the AC, as leading to TS, could have different perspectives* beyond the graphic defended by Kurzweil, and *his* expectations related to the road toward TS are problematic.

c. Is There an Explanation for Kurzweil's "Discoveries"?

One could ask: Why have Kurzweil and the defenders of AC reached the idea of exponential growth?

An appealing but unexpected answer is quite simple: they have searched for data to confirm their expectations.

As Michael Shermer observed, human beings are "pattern-seeking story-telling animals" [41].

This is why "we are quite adept at telling stories about patterns, whether they exist or not" [41].

He named this tendency "patternicity" [42].

Observing and evaluating, we are looking for and finding "patterns in our world and in our lives"; then, we "weave narratives around those patterns to bring them to life and give them meaning." Michael Shermer concludes: "such is the stuff of which myth, religion, history, and science are made" [43].

In Kurzweil's "discoveries," we have both *a subjective extension of an expectation*—AC, *and an alteration of the data in order to fit the model "discovered"*—the exponential growth of the returns toward the Singularity point.

Kurzweil's LAR is an example of connection created by its own expectations.

Yet, as one of the reviewers underlined, Kurzweil is neither the one who discovered TS nor the only one who is promoting it. However, again, due to his "eloquence," he is exemplary for the "infomorphic" paradigm. While his assumption remains controversial, even some clever works—such as Valentin F. Turchin's [44], mentioned by one of the reviewers of this paper—found a cybernetic approach in human evolution [44], while others promoted an entire cybernetic philosophy—as was the case with Mihai Draganescu's works [45–47], or have even questioned if we are living in a computer simulation [48]. One example of this flourishing debate and controversy is Ken Wharton's paper dismissing the computer simulation theory [49], or Zohar Ringel and Dmitry L. Kovrizhin's [50] or Andrew Masterson's conclusions on the same subject [51].

Those "infomorphic" paradigm-related theories of everything are, very probably, just examples of the (false) perceptions (and beliefs) created by our expectations.

As already underlined above, the status of AC as an objective tendency is still under debate. So is the status of LAR.

These are perceptions (and beliefs) created by our expectations—examples of patternicity. *AC (and LAR) are just examples of a specific patternicity case:* apophenia.

Apophenia is defined by the *Merriam-Webster Dictionary* as "the tendency to perceive a connection or meaningful pattern between unrelated or random things (such as objects or ideas)" [2], by the RationalWiki as "the experience of seeing meaningful patterns or connections in random or meaningless data" [52] and by *The Skeptic's Dictionary* as "the spontaneous perception of connections and meaningfulness of unrelated phenomena" [53].

Until now, several types of *apophenia* have been studied: *clustering illusion* ("the cognitive bias of seeing a pattern in what is actually a random sequence of numbers or events" [54]); *confirmation bias* ("the tendency for people to (consciously or unconsciously) seek out information that conforms to their pre-existing view points, and subsequently ignore information that goes against them, both positive and negative" [55]); *gambler's fallacy* ("the logical fallacy that a random process becomes less random, and more predictable, as it is repeated" [56]); and *pareidolia* ("the phenomenon of recognizing patterns, shapes, and familiar objects in a vague and sometimes random stimulus" [57]).

AC and LAR seem to be just cognitive biases related to the representation of future-related expectations.

4. There Is More than *Apophenia* in Kurzweil's TS; It Is *Pareidolia*

My two hypotheses about Kurzweil's famous "law of accelerating returns" (LAR) as undoubtedly leading to TS are the following.

1. *LAR is more likely just a new case of apophenia* [58,59]—as it shows "the spontaneous perception of connections and meaningfulness of unrelated phenomena" [53] and for centuries people have been perceiving the changes in science and technology as accelerating [23,58,60].

One of the reviewers of this paper wrote, "one absolutely cannot agree that exponential growth is a false pattern observed in random data as supposed by the notion of apophenia."

This is Kurzweil's opinion, too.

However, it is a false pattern not only because he manipulated data [23], but also, and more importantly, because exponential growth is just one of the possible models of growth [61] and it cannot continue indefinitely, but sometimes makes an+ inflexion and becomes an exponential decay [61,62] or just a slowdown [38].

The models of growth could have several types of representation, not only the exponential one.

"There is a whole hierarchy of conceivable growth rates that are slower than exponential and faster than linear (in the long run)" [61] and "growth rates may also be faster than exponential." In extreme cased, "when growth increases without bound in finite time, it is called hyperbolic growth. In between exponential and hyperbolic growth lie more classes of growth behavior" [38,61]. Ideally, growth continues "without bound in finite time" [38,61]. Sometimes exponential growth is simply slowdown [38,63].

When rejecting exponential growth as a most likely false pattern, we come up against the following problem: it is based on the evolutionary acquisition of patternicity as a specific human adaptive behavior, in order to ensure or facilitate individual or collective survival.

Indeed, "the search for pertinent patterns in the world is ubiquitous among animals, is one of the main brain tasks and is crucial for survival and reproduction." In the meantime, "it leads to the occurrence of false positives, known as patternicity: the general tendency to find meaningful/familiar patterns in meaningless noise or suggestive cluster" [64].

When claiming AC and consequently LAR and TS are objective tendencies, we are assuming everything is eventually explainable in a Mendeleevian-like table, in a solid, monolithic, and somehow mechanical explanation—so every other possibility should be rejected. Or, the world, human society, human beings, and very evolution of science and technology are complex and hardly predictable, as demonstrated in "The Age of Surprise" report [6].

Claiming AC/LAR/TS are objective tendencies leads, necessarily, to a *subjective* selection of data *considered trustworthy because it fits our expectations*. Or this is a new form of apophenia.

So, how can we trust the claims related to TS's possibility or even inevitability? Under these circumstances, as one of the reviewers correctly observed, when, maybe, "we can easily trust the claims related to TS's possibility," "we cannot so easily trust the claims related to TS's inevitability." I would add here: our trust is also a patternicity result.

2. TS could be considered more likely as a new case of pareidolia [41], because TS is AGI-based, and AGI is commonly and uncritically understood as a human-like intelligence.

The specificity of LAR's apophenia and TS's pareidolia (through AGI) is related to the direction of our perceptions and expectations—they are both future-related.

The arguments for such claims will be deployed in the following pages.

a. The Perfect New World of TS

As we saw, Kurzweil's expectations (and beliefs) are the following: "technological change is exponential, contrary to the common-sense 'intuitive linear' view"; the "returns" are increasing exponentially; there is "exponential growth in the rate of exponential growth"; machine intelligence will surpass, within just a few decades, human intelligence, "leading to The Singularity—technological change so rapid and profound it represents a rupture in the fabric of human history" based on "the merger of biological and nonbiological intelligence, immortal software-based humans, and ultra-high levels of intelligence that expand outward in the universe at the speed of light" [19].

For Kurzweil and the Singularitarians [65]—the adepts, the defenders and those who promote Singularitarianism [66] in almost a religious way [67,68]—these expectations (and beliefs) seem to be confirmed by the pace of progress in science and technology.

Or, rather, there is no one pace of progress, but paces of progress.

Some critics of LAR observed that there are not only different rates of AC in technical innovation and scientific discovery, but also very different phenomena and processes appropriate to be mathematically included in graphical representation(s) of the accelerating change [38].

Meanwhile, for other researchers, TS is just intellectual fraud [69].

From such a perspective, Kurzweil's LAR and Richard W. Paul's plead for AC are just unnecessary reductions of the complexity of the tendencies in the evolution of science and technology, to the linearity and predictability of our expectations.

Unifying and reducing all the rhythms and paces of progress to one perspective and equation is an exercise of the imagination, under the question: How will the perfect future world be?

b. (Again) "Accelerating Change" (AC) as Apophenia

Considering *apophenia* and *pareidolia*, let us remember, once more, some of their characteristics.

They can occur simultaneously, observes Robert Todd Carroll, as in the case of "seeing a birthmark pattern on a goat as the Arabic word for Allah and thinking you've received a message from God" or, as when seeing not only "the Virgin Mary in tree bark but believing the appearance is a divine sign" [58]. Here he discovers both *apophenia* and *pareidolia*.

Yet, "seeing an alien spaceship in a pattern of lights in the sky is an example of pareidolia," but it becomes apophenia if you believe the aliens have picked you as their special envoy [62].

Moreover, continues Carroll, commonly, "apophenia provides a psychological explanation for many delusions based on sense perception"—"UFO sightings"; "hearing of sinister messages on records played backwards"—whereas "pareidolia explains Elvis, Bigfoot, and Loch Ness Monster sightings" [58].

Seeing the pattern of exponential acceleration in the pace of technological change and representing the exponential growth as an asymptote is apophenia, because there is no general consensus about the objective existence of AC (and LAR) and, in an attempt to break the ultimate unpredictability of the complexity of future evolution of technology and science, unrelated and/or unclearly related phenomena and/or processes are connected and considered meaningful based on our profound need for order [70].

It is a semiotic situation as "any fact becomes important when it's connected to another" (Eco).

Even though it is a model that seems to work for particular cases, it still proves the ultimate weakness of a reductionist approach, as was true with the Ptolemaic model, too. Let us remember that from a false statement we can reach both a true or a false conclusion—in this case, from a reductionist subjective model we could have both falsehood and truth in the idea of change in the evolution of technology. Because of this truth-falsehood status of the observation-based idea, there is sometimes accelerating change in some technologies' evolution; we cannot extrapolate to AC (as a general objective tendency) and consider the model we use—based on a positivist assumption of the technology's progress—to be necessarily true.

It is a special case of *patternicity*, an apophenia, when, based on subjective selection and *arbitrary inferences* [71], various evolutionary tendencies, from various fields, are merged into a single evolutionary pattern.

The graphic representation Kurzweil used to illustrate the growth of computing power is Figure 2 (above).

The data related to technological change and advancement have been altered, manipulated, and adjusted in order to fit his expectations related to AC, LAR, and TS.

Observing Kurzweil's "methodology," Nathan Pensky concluded, "Thus, 'evolution' can mean whatever Kurzweil wants it to mean." It requires joining "disparate types of 'evolution'" [72].

Under these conditions, "the graph takes an exponential curve not because humans have moved inexorably along a track of "accelerating returns," but because Kurzweil has "ordered" data points in order to "reflect the narrative he likes" [72].

The future-related TS's *apophenia* [73] is a good example of the *intentionality fallacy* described by David Christopher Lane [74] and explained by Sandra L. Hubscher in her article on *apophenia* from *The Skeptic's Dictionary* [59].

c. TS (through AGI) as Pareidolia

As a new type of *pareidolia*, TS could be described as a wrong, subjective visual representation of the expectation related to human future(s), under a human-like AGI presupposition/assumption, also using *arbitrary inferences* [71].

The best example is this common and uncritical expectation: AGI, leading, inexorable, to TS will be human-like—at least in its first stages.

As in this *Information* special issue, serious arguments have been deployed against such an idea [75].

Let us illustrate the enthusiastic defense of human-like AGI based on a short survey of bombastic headlines in the media: "One step closer to having human emotions: Scientists teach computers to feel 'regret' to improve their performance [76], "Daydreaming simulated by computer model" [77], "Kimera Systems Delivers Nigel—World's First Artificial General Intelligence" [78], "Meet the world's 'first psychopath AI'" [79], "Human-like A.I. will emerge in 5 to 10 years, say experts" [80], etc.

The expectation that AI/AGI is/will be human-like or will be congruent with human intelligence(s) and emotion(s) is the ground for a narcissistic and future-related TS pareidolia. In fact, AI/AGI could have different characteristics than human intelligence, even though we human beings have created the so-called "intelligent software" [81].

One cause of the anthropocentric claims of the Singularitarians could be this: we are often forgetting that, from a false assumption—in this case, there is AC/LAR and AGI will be, necessarily, a human-like intelligence—based on an apophenic/pareidolic future related hope, one can deduce anything.

An example is Roman Yampolskyi's rejection of Toby Walsh's idea that the Singularity may not be near [82], in this special issue of *Information*.

TS's pareidolia cannot be validated as true or false, even by the most brilliant minds.

5. The TS-Related "Complexity Fallacy"

Another cause of such anthropocentric claims could be related to a wrong understanding and management of complexity, not only in the research on TS, but in the very way we are creating our hardware and software.

All these discussions, debates, and controversies related to AC, AI/AGI, or TS seem to be like the birds' uncoordinated songs in a wood and not like a symphony. There is a rise in different perspectives, definitions, and claims related to them.

Under these circumstances, systematizations and classifications have been proposed in papers and/or books/collected papers signed/edited by various researchers: Anders Sandberg [31], Nick Bostrom [83], Nikita Danaylov (Socrates) [84], Amnon H. Eden, James H. Moor, Johnny H. Søraker and Eric Steinhart [32], Eliezer S. Yudkowsky [85], IEEE's *Spectrum. Special Report on Singularity* [86], John Brockman [87], Adriana Braga and Robert K. Logan [75], etc.

In fact, the history of science and technology is full of attempts to reduce the richness of the facts, phenomena, entities, and beings to a Mendeleevian-like table. "This is the Faustian knowledge management philosophy assumed by the Wizard Apprentice" [88].

This is "a sign of a deep belief in the power of the taxonomy." It is also "an effect of the so-called presupposition of the 'generic (=linear and fully predictable) universality'—one of the best expressions of a mechanistic perspective on the world." It is about "claiming that we could fully reverse a deduction," usually "through strait abduction, in an attempt to rebuild the so called unity of the unbroken original mirror of the human knowledge using its fragments" [88].

This is about dismissing complexity for the profit of reductionist simplicity.

I would name this reduction of what is complex—and so, nonlinear and unpredictable, but also partially predictable—to what is linear and predictable, when naturally predictable (just complicated), the complexity fallacy. There are many AC, LAR, AI/AFI, and TS approaches deploying such a fallacy as an effect of an unconscious *patternicity*.

This situation suggests that AC, LAR, AI/AGI, and TS have to be studied from a perspective really based on complexity, as was already suggested by Paul Allen—with his "complexity break" argument against TS [89] and Viorel Guliciuc—with his examples of differences between computers' and human networked minds' functioning [88].

Paull Allen observes, "the amazing intricacy of human cognition should serve as caution for those who claim that the Singularity is close," as "without having a scientifically deep understanding of the cognition, we cannot create the software that could spark the Singularity." So, "rather than the ever-accelerating advancement predicted by Kurzweil," it is more likely "that progress towards this understanding is fundamentally slowed by the complexity break" [89].

Complexity break is described in these words: "as we go deeper and deeper in our understanding of natural systems," we find we need "more and more specialized knowledge to characterize them, and we are forced to continuously expand our scientific theories in more and more complex ways." So, "understanding the detailed mechanisms of human cognition is a task that is subject to this complexity brake" [89].

As quoted in this special issue of *Information*, "human minds are incredibly complex" [76] and the way humans think in patterns is very different from AI/AGI data processing [90]. So, an AGI leading to TS should necessarily embody human-like emotions in cognition [91].

Moreover, AI researchers—and let us assume that the same is applicable to AGI and TS researchers—"are only just beginning to theorize about how to effectively model the complex phenomena that give human cognition its unique flexibility: uncertainty, contextual sensitivity, rules of thumb, self-reflection, and the flashes of insight that are essential to higher level thought" [89].

As Robert K. Logan and Adriana Braga argued in their essay on the weakness of the AGI hypothesis, there is real danger in devaluing "aspects of human intelligence" as one cannot ignore or consider in a reductionist way "imagination, aesthetics, altruism, creativity, and wisdom" [75].

There is no need here to consider again the discussion about the strong AGI's hypothesis and the dangers AGI's (human-like) misunderstanding could bring with it, already detailed in their paper and/or in other papers from this special issue of *Information* [90–92].

Instead, what it is important for this paper is the conclusion of one of the papers from this special issue that even id "it is possible to build a computer system that follows the same laws of thought and shows similar properties as the human mind," "such an AGI will have neither a human body nor human experience, it will not behave exactly like a human, nor will it be "smarter than a human" on all tasks" [93]. A similar conclusion related to something other than full human-like evolution of AGI is underlined by other researchers, too [81].

Accepting these observations, let us add the findings of Thomas W. Meeks and Dilip V. Jeste in the neurobiology of wisdom when dealing with uncertainty: "prosocial attitudes/behaviors, social decision making/pragmatic knowledge of life, emotional homeostasis, reflection/self-understanding, value relativism/tolerance" [94].

The AGI strong hypothesis is not just "very complicated," as noted by one of the reviewers of this paper, but complex. That is, complexity cannot be reduced to complicatedness.

However, the next observation of the reviewer can be fully accepted: "The author may want to revise the conclusion 'AGI is impossible' to 'The possibility of AGI cannot be established by the arguments provided via TS and AC'."

Indeed, we do not have, for now, enough evidence to decide if (human-like) AGI is possible or impossible, nor enough arguments to sustain the truth of the claim that AC, through AGI, will necessarily lead to TS.

This is why most of the current perspectives on AGI and TS seem to be unprepared to really deal with their complexities, and this is "why they are facing so many difficulties, uncertainties and so much haziness in the full and appropriate understanding of the TS" [88].

6. Conclusions

a. The appropriate study of AC, AI/AGI, and/or TS requires the complexity of networked minds in order to manage the complexity break and avoid the complexity fallacy and different forms of wrong patternicity.

The argument for such a claim is, somehow unexpectedly, offered by the very functioning of the social networks specialized and focused on research.

CrowdForge [95], EteRNA [96], and other experiments [97], for example, proved the power of networked minds having a research task, *when dealing with missing data*, to obtain, each time, "impossible" correct results, when the most powerful computers and software were not able to reach any correct result.

EteRNA players, for example, were "extremely good at designing RNA's." Their results were most surprising "because the top algorithms published by scientists are not nearly so good. The gap is pretty dramatic." This chasm was attributed to the fact that "humans are much better than machines at thinking through the scientific process and making intelligent decisions based on past results." This conclusion is of the greatest importance for AGI and/or TS studies as "a computer is completely flummoxed by not knowing the rules"; when human "players are unfazed: they look at the data, they make their designs, and they do phenomenally" [96].

What could explain the clear gap between human networked minds and computers' results?

I think the answer is this: our intelligent artifacts are built based on linear, predictable, and predictable reasoning and not based on complex, nonlinear, partially predictable, and unpredictable reasoning.

"Linear and predictable" in the above claim means without "imagination, aesthetics, altruism, creativity, and wisdom" [80]. Our intelligent artifacts are executing sets of logical steps—algorithms. They cannot imagine, create, feel, or be wise. Everything they do is measurable and predictable.

Human reasoning is so complex that it cannot be reduced to a single logical rule/type of reasoning or to any set of logical rules/types of reasoning covering all possibilities. Human reasoning is more than "complicated": it is complex and so *irreducible* to a machine-like model. In most cases, in human reasoning there is no unique logical rule compulsory to obtain a result—just because from a false claim/sentence/proposition we will always be in a position to obtain both the truth and the falsehood. Human logical "machines" have holes in their functionning. There is some predictability in human reasoning, but it remains *not fully predictable*.

Yet, when it comes to the future, and especially TS, considered as a "rupture in the fabric of human history" [19], we cannot have enough predictable information about how it would be because we do not know what, why, and how exactly TS will be or, even, more likely, *if* TS will be, so it is unpredictable.

For example, even the merging of humans with machines is complicated, as there are many meanings, types, and grades of merging [98].

So, any number of networked computers will retain this weakness: they cannot find a result if some data is missing, when social networked minds can—as in the examples above.

We have to keep in mind "the power of the human mind to collectively surpass the power of computation of our 'smartest' machines just because the machine (=AI/AGI), being created using a linear reasoning, cannot deal with the complexity" [88].

b. TS would require redesigning AGI based on complexity—which we are not sure is possible.

Reaching TS (through AGI) seems to not be possible without reaching real complexity (not complicatedness!) in designing our "intelligent" artifacts. Redesigning hardware-software systems based on nonlinearity and unpredictability is not yet possible without fully understanding the complexity of our human, not-machine, minds. Maybe it will never be possible. Until then, TS is more likely a creation of our best expectations, an example of pareidolia, based on reductionism, subjective extrapolation, and imagination.

So, let us think (digitally) wisely and wait for the surprises the future(s) is/are preparing for us already!

Funding: This research received no external funding.

Acknowledgments: The author expresses his deep gratitude for the reviewers' comments, for the editors of this special issue, for my friend Dr. Iulian Rusu – the Editior iu Chief of the *European Journal of Science and Theology*, for the assistant editor assigned for this paper submitted to *Information,* for the copy-editors, proofreaders and typesetting specialists of MDPI, and to his wife MA Gabriela Guliciuc. The author kindly asks that no reference to his ideas deployed in this paper should be in any way used without his permission until the publishing of this paper in *Information* or another journal.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Wiktionary. Patternicity. Available online: https://en.wiktionary.org/wiki/patternicity (accessed on 21 October 2018).
2. Meriam-Webster Dictionary. Apophenia. Available online: https://www.merriam-webster.com/dictionary/apophenia (accessed on 21 October 2018).
3. Meriam-Webster Dictionary. Pareidolia. Available online: https://www.merriam-webster.com/dictionary/pareidolia (accessed on 21 October 2018).
4. Aristotle. *Metaphysics, Book 9.8*; Ross, W.D., Translator; O.U.P., American Branch; 1924; pp. 91–92. Available online: http://www.documentacatholicaomnia.eu/03d/-384_-322,_Aristoteles,_13_Metaphysics,_EN.pdf (accessed on 21 October 2018).
5. Cohen, R. The Age of Surprise: Predicting the Future of Technology. *Forbes*. 18 December 2013. Available online: https://www.forbes.com/sites/reuvencohen/2013/12/18/the-age-of-surprise-predicting-the-future-of-technology/#275f606f6570 (accessed on 31 December 2013).
6. Air University. USAF Center for Strategy and Technology. Wellcome to 2035 ... the Age of Surprise. 8 September 2012. Available online: https://www.airuniversity.af.mil/CSAT/ (accessed on 31 December 2013).
7. Bostrom, N.; Ćirković, M.M. (Eds.) *Global Catastrophic Risks*; Oxford University Press: Oxford, UK, 2011.
8. Wikiversity. Introduction to Futures Studies. Available online: https://en.wikiversity.org/wiki/Introduction_to_Futures_Studies (accessed on 31 October 2017).
9. World Futures Studies Federation, About Futures Studies. Available online: https://www.wfsf.org/about-us/futures-studies (accessed on 3 September 2018).
10. Lassiter, D.; Goodman, N.D. How Many Kinds of Reasoning? Inference, Probability, and Natural Language Semantics. *Cognition* **2017**, *136*, 123–134. [CrossRef] [PubMed]

11. Arbesman, S. *The Half-life of Facts: Why Everything We Know Has an Expiration Date*; Penguin Group: London, UK, 2012; ISBN 978-1591844723.
12. Stein, R.M. The Half-life of Facts: Why Everything We Know Has an Expiration Date. *Quant. Financ. Themed Issue Deriv. Pricing Hedging* **2014**, *14*, 1701–1703. [CrossRef]
13. Evans, J.A. Future Science. *Science* **2013**, *342*, 44–45. [CrossRef] [PubMed]
14. Aristotle. *On Interpretation*; University of Adelaide: eBooks@Adelaide; Edghill, E.M., Translator; 2015. Available online: https://ebooks.adelaide.edu.au/a/aristotle/interpretation/ (accessed on 21 October 2018).
15. Stanford Encyclopedia of Philosophy. Future Contingents. Available online: https://plato.stanford.edu/entries/future-contingents/ (accessed on 20 October 2017).
16. Charles Sanders Peirce. *Collected Papers of Charles Sanders Peirce*; The Belknap Press: Cambridge, MA, USA, 1958; Volume 7.
17. Nedelescu, L. The Rising Toll of the (Still) Predominant Mechanistic Mindset in a Complex World. 11 June 2013. Available online: https://generative-management.com/2013/06/11/the-insurmountable-toll-of-the-still-predominantly-mechanistic-mindset-in-a-complex-world/ (accessed on 15 August 2018).
18. Smart, J.M.; Foresight, U. FERN. The Foresight Guide. Predicting, Creating, and Leading in the 21st Century (Alpha Version. Chapters 7, 8, and 10 Still Being Written). Available online: http://www.foresightguide.com/universal-accelerating-change/ (accessed on 12 March 2018).
19. Kurzweil, R. The Law of Accelerating Returns. 7 March 2001. Available online: http://www.kurzweilai.net/the-law-of-accelerating-returns (accessed on 10 September 2017).
20. Eliazar, I.; Shlesinger, M.F. Universality of accelerating change. *Phys. A Stat. Mech. Appl.* **2018**, *494*, 430–445. [CrossRef]
21. Paul, R.W. *How to Prepare Students for a Rapidly Changing World*; Foundation for Critical Thinking: Dillon Beach, CA, USA, 1995. Available online: http://www.criticalthinking.org/pages/accelerating-change/474 (accessed on 11 September 2017).
22. Martin, B.R. The origins of the concept of 'foresight' in science and technology: An insider's perspective. *Technol. Forecast. Soc. Chang.* **2010**, *77*, 1438–1447. [CrossRef]
23. Modis, T. Why the Singularity Cannot Happen. In *Singularity Hypotheses. A Scientific and Philosophical Assessment*; Eden, A.H., Moor, J.H., Søraker, J.H., Steinhart, E., Eds.; Springer Verlag: Berlin, Germany, 2012; pp. 311–339.
24. Huebner, J. A possible declining trend for worldwide innovation. *Technol. Forecast. Soc. Chang.* **2005**, *72*, 980–986. [CrossRef]
25. Korotayev, A.; Malkov, A.; Khaltourina, D. *Introduction to Social Macrodynamics: Secular Cycles and Millennial Trends*; Editorial URSS: Moscow, Russia, 2006; ISBN 5-484-00559-0. Available online: https://www.researchgate.net/profile/A_Korotayev/publication/233821695_Introduction_to_social_macrodynamics_Secular_cycles_and_millennial_trends/links/0912f50c1fb3ea509e000000/Introduction-to-social-macrodynamics-Secular-cycles-and-millennial-trends.pdf (accessed on 10 November 2018).
26. Lane, J. Assessing the Impact of Science Funding. *Science* **2009**, *324*, 1273–1275. [CrossRef] [PubMed]
27. Jones, R. Accelerating Change or Innovation Stagnation? *Soft Machines*. 25 March 2011. Available online: http://www.softmachines.org/wordpress/?p=1027 (accessed on 22 October 2018).
28. Moschella, D. The Pace of Technology Change is Not Accelerating. *Leading Edge Forum*. 2 September 2015. Available online: https://leadingedgeforum.com/publication/the-pace-of-technology-change-is-not-accelerating-2502/ (accessed on 22 October 2018).
29. Wikipedia. Accelerating Change. Available online: https://en.wikipedia.org/wiki/Accelerating_change (accessed on 20 January 2018).
30. Smart, J. A Taxonomy of Singularities: Comparing the Literature on Systems of Accelerating Change. *Acceleration Watch*. Available online: https://www.accelerationwatch.com/taxonomyofsingularities.html (accessed on 10 November 2018).
31. Sandberg, A. An overview of models of technological Singularity. In *The Transhumanist Reader: Classical and Contemporary Essays on the Science, Technology, and Philosophy of the Human Future*; More, M., Vita-More, N., Eds.; Wiley Blackwell: Hoboken, NJ, USA, 2013; pp. 376–394. Available online: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118555927.ch36 (accessed on 3 June 2018).

32. Eden, A.H.; Moor, J.H.; Søraker, J.H.; Steinhart, E. (Eds.) *Singularity Hypotheses. A Scientific and Philosophical Assessment*; Springer Verlag: Berlin, Germany, 2012.
33. Kurzweil, R. *The Singularity Is Near. When Humans Transcend Biology*; Viking, Penguin Group: New York, NY, USA, 2005. Available online: http://stargate.inf.elte.hu/~{}seci/fun/Kurzweil,%20Ray%20-%20Singularity%20Is%20Near,%20The%20%28hardback%20ed%29%20%5Bv1.3%5D.pdf (accessed on 11 February 2017).
34. Wichmann, J. Our World Is Changing—But Not As Rapidly As People Think. 2 August 2018. Available online: https://www.weforum.org/agenda/2018/08/change-is-not-accelerating-and-why-boring-companies-will-win/ (accessed on 20 January 2018).
35. Magee, C.L.; Devezas, T.C. How many singularities are near and how will they disrupt human history? *Technol. Forecast. Soc. Chang.* **2011**, *78*, 1365–1378. [CrossRef]
36. Modis, T. The Singularity Myth. *Technol. Forecast. Soc. Chang.* **2006**, *73*. Available online: https://www.researchgate.net/publication/267207324_The_Singularity_Myth (accessed on 10 September 2017).
37. TutorVista. Asymptotes of Rational Functions. Available online: https://math.tutorvista.com/calculus/asymptotes-of-rational-functions.html (accessed on 16 April 2018).
38. Korotayev, A. The 21st Century Singularity and its Big History Implications: A re-analysis. *J. Big Hist.* **2018**, *2*, 73–119. [CrossRef]
39. Underground Mathematics. Approaching Asymptotes. Available online: https://undergroundmathematics.org/thinking-about-functions/approaching-asymptotes/things-you-might-have-noticed (accessed on 16 April 2018).
40. Simonite, T. Moore's Law Is Dead. Now What? *MIT Technology Review*. 13 May 2016. Available online: https://www.technologyreview.com/s/601441/moores-law-is-dead-now-what/ (accessed on 11 November 2018).
41. Shermer, M. Out of This World. *The Washington Post*. 21 November 1999. Available online: https://www.washingtonpost.com/wp-srv/WPcap/1999-11/21/005r-112199-idx.html??noredirect=on (accessed on 8 May 2017).
42. Shermer, M. Patternicity: Finding Meaningful Patterns in Meaningless Noise. Why the Brain Believes Something Is Real When It Is Not. *Scientific American*. 1 December 2008. Available online: https://www.scientificamerican.com/article/patternicity-finding-meaningful-patterns/ (accessed on 8 January 2018).
43. Shermer, M. Chicken Soup for the Evolutionist's Soul. A review of Robert Wright's Nonzero: The Logic of Human Destiny. 6 February 2000. Available online: https://michaelshermer.com/2000/02/chicken-soup-for-the-evolutions-soul/ (accessed on 15 September 2018).
44. Turchin, V.F. *The Phenomenon of Science: A Cybernetic Approach to Human Evolution*; Columbia University Press: New York, NY, USA; Guildford, UK, 1977. Available online: http://pespmc1.vub.ac.be/POS/TurPOS-prev.pdf (accessed on 11 November 2018).
45. Draganescu, M. Profunzimile Lumii Material. Bucuresti: Editura Politica, 1979/The Depths of Existence, Online Edition. 1997. Available online: http://www.racai.ro/external/static/doe/ (accessed on 10 November 2013).
46. Draganescu, M. *Ortofizica—Încercare Asupra Lumii și Omului din Perspectiva Științei Contemporane*; Editura Științifică și Enciclopedică: București, Romania, 1985.
47. Draganescu, M. *Inelul Lumii Materiale*; Editura Științifică și Enciclopedică: București, Romania, 1989.
48. Moskowitz, C. Are We Living in a Computer Simulation? *Scientific American*. 7 April 2016. Available online: https://www.scientificamerican.com/article/are-we-living-in-a-computer-simulation/ (accessed on 12 November 2018).
49. Wharton, K. The Universe Is Not a Computer. 27 January 2015. Available online: https://arxiv.org/pdf/1211.7081v2 (accessed on 12 November 2018).
50. Ringel, Z.; Kovrizhin, D.L. Quantized gravitational responses, the sign problem, and quantum complexit. *Science* **2017**, *3*, e1701758. Available online: http://advances.sciencemag.org/content/3/9/e1701758 (accessed on 12 November 2018).
51. Masterson, A. Physicists Find We're Not Living in A Computer Simulation. 2 October 2017. Available online: https://cosmosmagazine.com/physics/physicists-find-we-re-not-living-in-a-computer-simulation (accessed on 12 November 2018).

52. RationalWiki. Apophenia. Available online: https://rationalwiki.org/wiki/Apophenia (accessed on 21 October 2018).
53. The Skeptic Dictionary. Apophenia. Available online: http://skepdic.com/apophenia.html (accessed on 21 October 2018).
54. RationalWiki. Clustering Illusion. Available online: https://rationalwiki.org/wiki/Clustering_illusion (accessed on 21 October 2018).
55. RationalWiki. Confirmation Bias. Available online: https://rationalwiki.org/wiki/Confirmation_bias (accessed on 21 October 2018).
56. RationalWiki. Gambler's Fallacy. Available online: https://rationalwiki.org/wiki/Gambler%27s_fallacy (accessed on 21 October 2018).
57. RationalWiki. Pareidolia. Available online: https://rationalwiki.org/wiki/Pareidolia (accessed on 21 October 2018).
58. Carroll, R.T. Apophenia. In *The Skeptic's Dictionary: A Collection of Strange Beliefs, Amusing Deceptions, and Dangerous Disillusions*; John Willey & Sons: Hoboken, NJ, USA, 2003. Available online: http://skepdic.com/apophenia.html (accessed on 11 May 2018).
59. Hubscher, S.L. Apophenia: Definition and Analysis. *Digital Bits Skeptic*. 4 November 2007. Available online: http://www.dbskeptic.com/2007/11/04/apophenia-definition-and-analysis/#selection-15.0-15.20 (accessed on 17 September 2018).
60. Burke, J.; Bergman, J.; Asimov, I. The Impact of Science on Society, the Impact of Science. NASA SP-482; 1985; p. 16. Available online: https://history.nasa.gov/sp482.pdf (accessed on 17 September 2018).
61. Wikipedia Exponential Growth. Available online: https://en.wikipedia.org/wiki/Exponential_growth (accessed on 2 October 2018).
62. Math Is Fun. Exponential Growth and Decay. Available online: https://www.mathsisfun.com/algebra/exponential-growth.html (accessed on 2 October 2018).
63. Boucher, D. The World's Population Hasn't Grown Exponentially for at Least Half a Century. *Union of Concerned Scientists*. 9 April 2018. Available online: https://blog.ucsusa.org/doug-boucher/world-population-growth-exponential (accessed on 2 October 2018).
64. Correa Varella, M.A. The Biology and Evolution of the Three Psychological Tendencies to Anthropomorphize Biology and Evolution. *Front. Psychol.* **2018**. [CrossRef]
65. Danailov, N. Top 10 Singularitarians of All Time, 23 January 2001. Available online: https://www.singularityweblog.com/top-10-singularitarians/ (accessed on 1 November 2017).
66. Wikipedia. Singularitarianism. Available online: https://en.wikipedia.org/wiki/Singularitarianism (accessed on 11 September 2018).
67. Horgan, J. The Consciousness Conundrum. *IEEE Spectrum's Special Report: The Singularity*. 1 June 2008. Available online: https://spectrum.ieee.org/biomedical/imaging/the-consciousness-conundrum/0 (accessed on 26 August 2017).
68. Guliciuc, V. Transhumanism—A New Faith or a New Religion? In Proceedings of the ESRARC 2018 10th European Symposium on Religious Art, Restoration & Conservation, Prague, Czech Republic, 31 May–1 June 2018; pp. 198–202.
69. Pein, C. The Singularity Is Not Near: The Intellectual Fraud of the "Singularitarians". 13 May 2018. Available online: https://www.salon.com/2018/05/13/the-singularity-is-not-near-the-intellectual-fraud-of-the-singularitarians/ (accessed on 10 September 2018).
70. Hale, J. Patterns: The Need for Order. *PsychCentral*. 17 July 2016. Available online: https://psychcentral.com/lib/patterns-the-need-for-order/ (accessed on 2 October 2018).
71. Tolboll, M. *A Dictionary of Thought Distortions*; WingSpan Press: Livermore, CA, USA, 2014; pp. 6–7.
72. Pensky, N. Ray Kurzweil Is Wrong: The Singularity Is Not Near. 3 February 2014. Available online: https://pando.com/2014/02/03/the-singularity-is-not-near/ (accessed on 12 June 2018).
73. Carroll, R.T. Apophenia and Pareidolia. *Unnatural Acts That Can Improve Your Thinking*. 9 January 2012. Available online: http://59ways.blogspot.com/2012/01/apophenia-and-pareidolia_09.html (accessed on 7 September 2018).
74. Lane, D.C.; Lane, A.-D. Apophenia and the Intentionality Fallacy. Why License Plates are Not Messages from the Beyond. December 2010. Available online: http://www.integralworld.net/lane17.html (accessed on 10 October 2017).

75. Braga, A.; Logan, R.K. The Emperor of Strong AI Has No Clothes: Limits to Artificial Intelligence. *Information* **2017**, *8*, 156. Available online: https://www.mdpi.com/2078-2489/8/4/156 (accessed on 12 September 2018).
76. Bates, D. One Step Closer to Having Human Emotions: Scientists Teach Computers to Feel 'Regret' to Improve Their Performance. *Mail Online*. 19 April 2011. Available online: https://www.dailymail.co.uk/sciencetech/article-1378464/Scientists-teach-computers-feel-regret-improve-performance.html (accessed on 17 September 2018).
77. Purdy, M.C. Daydreaming simulated by computer model. *The Source*. 11 July 2013. Available online: https://source.wustl.edu/2013/07/daydreaming-simulated-by-computer-model/ (accessed on 4 April 2017).
78. Busines Wire. Kimera Systems Delivers Nigel—World's First Artificial General Intelligence. 10 August 2016. Available online: https://www.businesswire.com/news/home/20160810005371/en/Kimera-Systems-Delivers-Nigel-%E2%80%93-World%E2%80%99s-Artificial (accessed on 11 August 2016).
79. McKenna, J. Meet the World's 'First Psychopath AI'. *World Economic Forum*. 13 June 2018. Available online: https://www.weforum.org/agenda/2018/06/scientists-create-world-first-psychopath-ai-norman/ (accessed on 21 June 2018).
80. Johnson, S.; Human-like A.I. Will Emerge in 5 to 10 Years, Say Experts. *Big Think*. 26 September 2018. Available online: https://bigthink.com/surprising-science/computers-smart-as-humans-5-years?rebelltitem=2#rebelltitem2 (accessed on 21 October 2018).
81. Montemayor, C. Human-Like Consciousness and Human-Like Intelligence. Human-like consciousness and human-like intelligence might evolve differently. *Psychology Today*. 28 August 2017. Available online: https://www.psychologytoday.com/us/blog/theory-consciousness/201708/human-consciousness-and-human-intelligence (accessed on 7 September 2018).
82. Yampolskiy, R.V. The Singularity May Be Near. *Information* **2018**, *9*, 190. [CrossRef]
83. Bostrom, N. Singularity and Predictability. *Extropy*. 1998. Available online: http://mason.gmu.edu/~{}rhanson/vc.html#bostrom (accessed on 1 September 2017).
84. Danaylov, N. 17 Definitions of the Technological Singularity. 18 April 2012. Available online: https://www.singularityweblog.com/17-definitions-of-the-technological-singularity (accessed on 11 June 2017).
85. Yudkowsky, E.S. Three Major Singularity Schools. 30 September 2007. Available online: https://intelligence.org/2007/09/30/three-major-singularity-schools/ (accessed on 7 September 2018).
86. IEEE Spectrum. Special Report: The Singularity. 2008. Available online: http://spectrum.ieee.org/static/Singularity (accessed on 1 June 2018).
87. Brockman, J. (Ed.) *What to Think About Machines that Think*; Harper Perennial: New York, NY, USA, 2015.
88. Guliciuc, V. Technological Singularity in the Age of Surprise facing complexity. *Eur. J. Sci. Theol.* **2014**, *10*, 79–88. Available online: http://www.ejst.tuiasi.ro/Files/46/8_Guliciuc.pdf (accessed on 13 October 2018).
89. Allen, P.G. The Singularity Isn't Near. *MIT Tecchnology Review*. 12 October 2011. Available online: https://www.technologyreview.com/s/425733/paul-allen-the-singularity-isnt-near (accessed on 1 September 2018).
90. Logan, R.K.; Tandoc, M. Thinking in Patterns and the Pattern of Human Thought as Contrasted with AI Data Processing. *Information* **2018**, *9*, 83. [CrossRef]
91. Lunceford, B. Love, Emotion and the Singularity. *Information* **2018**, *9*, 221. [CrossRef]
92. Baum, S.D. Countering Superintelligence Misinformation. *Information* **2018**, *9*, 244. [CrossRef]
93. Wang, P.; Liu, K.; Dougherty, Q. Conceptions of Artificial Intelligence and Singularity. *Information* **2018**, *9*, 79. [CrossRef]
94. Meeks, T.W.; Jeste, D.V. Neurobiology of wisdom: A literature overview. *Arch Gen Psychiatry* **2009**, *66*, 355–365. [CrossRef] [PubMed]
95. Rae, T. Carnegie Mellon Researchers Find Crowds Can Write as Well as Individuals. *Wired Campus -The Chronicle of Higher Education*. 3 February 2011. Available online: https://www.chronicle.com/blogs/wiredcampus/carnegie-mellon-researchers-find-crowds-can-write-as-well-as-individuals/29440 (accessed on 17 September 2018).
96. Wiseman, R. The Public, Playing A Molecule-Building Game, Outperforms Scientists. *Wired Campus-The Chronicle of Higher Education*. 12 August 2011. Available online: https://www.chronicle.com/blogs/wiredcampus/the-public-playing-a-molecule-building-game-outperforms-scientists/32835 (accessed on 17 September 2018).

97. Young, J.R. Crowd Science Reaches New Heights. *Technology -The Chronicle of Higher Education*. 28 May 2010. Available online: https://www.chronicle.com/article/The-Rise-of-Crowd-Science/65707 (accessed on 17 September 2018).
98. Guliciuc, V. From Wisdom to Digital Wisdom as Negotiated Identity. *Eur. J. Sci. Theol.* **2013**, *9*, 1–15.

information

MDPI

Article

Cosmic Evolutionary Philosophy and a Dialectical Approach to Technological Singularity

Cadell Last

Evolution, Cognition, and Complexity (ECCO) group, Global Brain Institute; B-1160 Brussels, Belgium; cadell.last@gmail.com

Received: 13 February 2018; Accepted: 4 April 2018; Published: 5 April 2018

Abstract: The anticipated next stage of human organization is often described by futurists as a global technological singularity. This next stage of complex organization is hypothesized to be actualized by scientific-technic knowledge networks. However, the general consequences of this process for the meaning of human existence are unknown. Here, it is argued that cosmic evolutionary philosophy is a useful worldview for grounding an understanding of the potential nature of this futures event. In the cosmic evolutionary philosophy, reality is conceptualized locally as a universal dynamic of emergent evolving relations. This universal dynamic is structured by a singular astrophysical origin and an organizational progress from sub-atomic particles to global civilization mediated by qualitative phase transitions. From this theoretical ground, we attempt to understand the next stage of universal dynamics in terms of the motion of general ideation attempting to actualize higher unity. In this way, we approach technological singularity dialectically as an event caused by ideational transformations and mediated by an emergent intersubjective objectivity. From these speculations, a historically-engaged perspective on the nature of human consciousness is articulated where the truth of reality as an emergent unity depends on the collective action of a multiplicity of human observers.

Keywords: futures; technological singularity; philosophy; cosmic evolution; anthropology

1. Global Technological Singularity

Scientific networks constructed by human psychosocial processes have made measurable progress in understanding the natural world analytically through the tools of empirical observation and the methods of reduction and fragmentation. In these efforts, reduction refers to the practice of understanding nature by isolating particular phenomena in nature and analyzing its constituent parts like particles, molecules, and organisms [1]. The logic of reductionism has led to fragmentation of natural analysis into fields and sub-fields and sub-sub-fields focusing on ever more specified parts within the wider whole like physics, chemistry, and biology [2]. This dynamic analytic process is necessary for certain forms of understanding nature. For example, fields as diverse as M-theory in particle physics [3], or artificial intelligence in cognitive science [4], or genetic engineering in biology [5], could not exist without the tools of empirical observation becoming channeled through methods of reduction and fragmentation.

However, there is an emerging intellectual desire and social necessity in the philosophy of science and in the sciences of humanity to understand the historical dialectic consequences of scientific networks in relation to nature and humanity as a totality. This desire and necessity is due to an increasingly unpredictable, uncertain, and chaotic future horizon of becoming [6]. In short, scientific reduction and fragmentation allow us to understand diverse objective phenomena, but is unable to help us understand the emergent holistic consequence of this understanding for subjectivity. For example, throughout the history of modern science the human existential position and relation to nature has become de-centered from the classical philosophical immediacy of being [7], and also the traditional religious affirmation of God [8]. These de-centerings are often framed in cosmic terms (e.g., Copernican heliocentrism), biological terms (e.g., Darwinian selectionism), and psychological terms (e.g., Freudian unconscious) [9]. Such symbolic movements generated by scientific knowledge leave us devoid of absolute value, consequently producing fundamental metaphysical crises of human meaning vis-à-vis being itself [10].

In short, metaphysical de-centerings caused by scientific networks have allowed for deeper reductionist understanding of nature, like understanding the nature of solar systems and galaxies, the nature of evolution of living forms, and the nature of unconscious dreams and desires. However, this knowledge has simultaneously resulted in a broken and incoherent visions of holistic totality in relation to classical or traditional metaphysics. If we define humanity by the peculiar nature of our self-consciousness (awareness of being, reflection on being), then philosophical understandings of totality situated humanity in relation to either ideality in a transcendental superspace or a historical superspace as attractor [11]. However, in our contemporary age dominant conceptions of totality often become grounded in scientific materialist conceptions of cosmic or sub-atomic scale. In these visions of totality self-consciousness appears without experiential and transcendent aims of higher significance. We are frequently presented with narratives suggesting that the cosmos only has an inhuman aim of heat death leading to universal void, that biology has the inhuman aim of fitness maximization leading to endless living form replication, and that our own minds have a mysterious inhuman aim within related to unconscious drives and repetitions. From this understanding of totality notions of transcendent or historical ideality become difficult to convincingly substantiate on the terms of the critique of reason.

In this way the closed and complete worlds of classical or traditional metaphysics structured around the central importance of our self-conscious experience (on Earth) and its connection with divine transcendence (in Heaven) have been severed with the thought tools of Cartesian doubt and Kantian criticism [12]. The Cartesian cogito introduced the world to the centrality of rational thought as self-certainty of being, and Kantian transcendentals introduced the world to the a priori frame structuring being. Both philosophical tools prevent any connection in real knowledge to a closed and complete 'other world' capable of absolutely centering our being. This gives way for the growing chaos and uncertainty of the open and incomplete worlds of modern rationalism, where the central place of a fundamental ideal truth for human beings has no place outside of its own ego-centric bubble of illusion. In this frame ultimate or absolute existential meaning is replaced with a type of ultimate or absolute nihilism of being surrounded on all sides by forces working against our own interests. However, this situation grows more complex when we consider the mystery of the consequences of scientific epistemology in its broadest context of universal history [13]. To be specific the activity of scientific networks grounded in conceptual reduction and fragmentation can be logically extrapolated towards a futures horizon of a global technological singularity [14].

The notion of technological singularity driven by scientific networks actualizing artificial intelligence, genetic engineering, and quantum computing (for example) challenges theorists to think a fundamental impossibility of thought. This impossibility for thought is ultimately a qualitative phase transition representing a level of existence beyond our mental or ideational capability to meaningfully abstract. Thus, from the perspective of the human mind the technological singularity represents a type of infinity since it could involve the emergence of a phenomenal realm of higher level perception, thought, and communication, potentially in higher dimensional geometries than the three dimensional geometries (plus time) that fundamentally characterize and constrain human experience and aims. In terms of analytic mechanisms the most frequently deployed conjectures for the actual realization of such an event include artificial general intelligence (AGI) or human intelligence amplification (IA) [15–17].

In terms of scientific theory predictions for the emergence of a technological singularity are based on extrapolation models of exponential computational evolution. These models were originally founded on Moore's Law [18,19], and are currently often based on the 'Law of Accelerating Returns'

(LOAR) [20,21]. Moore's law and LOAR suggest that in terms of pure computational capacity the total cognitive capacities of the human species will be surpassed before mid-century. This is why extrapolations of this process are represented with attractive metaphors from mathematics and physics of an approaching 'event horizon' towards an 'other side' of being [22]. Currently there is no coherent scientific model that helps us to describe the actualization of embodied superintelligence (in whatever form it takes) and the ontological effects of its higher order cognitive force. In other words, what will a civilization with such high levels of computing power be actualizing and how can we theoretically approach such phenomena? [23].

Thus the futures horizon of technological singularity signifies an immanent actuality of a post-human world where our own modernist social systems structured by reason and science give rise to a suite of technological possibilities that undermine both human existence and transcendence [24]. From the proper historical context the technological singularity may well represent yet another scientific de-centering where humanity's fundamental social and historical mode of being itself appears with an uncontrollable inhuman aim immanent to our nature [25,26]. In this frame social systems as symbolic networks of communication, here in the form of reason and science, overdetermine the highest experiential and emotional interests of humans themselves (self-consciousness) in favor of their own unbounded and unconscious propagation [27].

In popular discourse these issues of our uncontrollable social and historical modes of being are generally framed as part of the Anthropocene [28]. The Anthropocene signifies the fact that scientific epistemology and general human presence is both increasingly powerful and increasingly unpredictable in its ontological consequences [29]. However, in this context technologically mediated issues of global warming and economic inequality appear insignificant in comparison to the possibility of post-human systems of machinic superintelligence [21]. Are these processes evolutionarily determined to eliminate possibilities of human experience and transcendence [30,31], or is it possible to gain control of these social and historical modes of being and direct them towards actualizing desires of concern for human experience and emotion? [32,33] In either context, how can we make sense of totality in relation to the truth of nature and humanity as an integrated whole? (Table 1).

Table 1. Potential large-scale singularity consequences for self-consciousness. The radical unknown and unpredictable future horizon of human becoming in relationship to scientific advancements in robotics, artificial intelligence, nanotechnology, genetics, and other possible actualizations is becoming a growing and central concern of human evolution, possibly requiring dialectical analysis. In this dialectical analysis the internal horizon of negative-affirmative ideality structures becoming around a center singularity.

Pathway.	Phenomena
Artificial Intelligence (AI) Scenario	Super-intelligent/conscious technological systems replace biocultural humans
Intelligence Amplification (IA) Scenario	Biocultural humans transform their being with technological systems becoming trans/post-humans
AI-IA Scenario	Complex interrelation of both organizational processes occur

When thinking about totality inclusive of a possible technological singularity we must first situate it in its historical ideality because its archetypal structure is not distinct from many other ideational structures internal to the symbolic order. In fact, the singularity, in terms of its attractive intensities and qualities of higher mentality, may be conceptualized as a techno-scientific representation that recaptures the metaphorical structure of many foundational religious traditions [34]. In particular this techno-scientific singularity thought structure resembles the thought structure of Western Abrahamic traditions like Judaism, Christianity, and Islam which formed the metaphysical ground for historical civilization before scientific modernity [35]. For Western religious traditions fundamental metaphysics natural being as given to self-consciousness is conceived of as insufficient or lacking due to fundamental experiential coordinates of morality, finitude, suffering, etc., and can only be remedied or reconciled by a future qualitative phase transition that restructures the relation between natural being and self-consciousness. Of course, in predominantly ancient religious narratives and value structures such a qualitative phase transition occurs via death or resurrection where the disembodied soul escapes the constraints of the world and enters into the sublime gates of an eternal ideality destined to enjoy spiritual immortality with God [36].

Here the principal difference between the historical religious conceptions of metaphysical singularity and the technological singularity is that technological singularity theorists ground their approach towards an immanent qualitative phase transition for thought in both secular and evolutionary terms that can ultimately be traced to the origin of scientific thought itself [37,38]. There are two frequently deployed analogies from cosmic evolution to describe this future. The first analogy is the transition between simple chemical processes and complex cellular biology. This process in chemistry and biology is referred to as abiogenesis [39], and is often associated with the emergence of life, and sometimes also mind [40]. The second analogy is the transition between biological great apes and biocultural human tribes. This process in anthropology is often associated with the emergence of self-consciousness and language [41], and in theology and philosophy often associated with the spiritualization of the cosmos with a telos repetitively aiming for reconciliation of natural being [42,43]. These past secular events of the emergence of biology and language are evoked as relevant for future cognitive process in order to communicate the fact that the emergence of a previously non-existent evolutionary potentiality has actually occurred before over big historical time [44].

2. Cosmic Evolutionary Philosophy

In order to approach questions of totality today perhaps we should first consider utilizing the philosophical worldview of cosmic evolution. This worldview has been most frequently deployed for the tasks of both articulating notions of technological singularity [45] and situating human development in a cosmic context [46]. The cosmic evolutionary worldview attempts to build a philosophy that can approach an integrated holistic view of totality as opposed to the fragmented reductionist view [47] (Figure 1). Thus, in contrast to the approaches of reduction and fragmentation the cosmic evolutionary approach focuses on a holistic integration of all phenomenon as unified in which totality is analyzed from the beginning of known processual dynamics to the present moment [45,48].

In this view we cannot understand totality without explaining the evolution of all networked phenomena on all scales of reality [49]. For example, instead of separating the world of physics which focuses on particle interactions and fields of force, and the world of chemistry which focuses on chemical interactions and autocatalytic cycles; a holistic and integrated view focuses on understanding the shared dynamic processes that connect the world of physics and the world of chemistry in an emergent multi-level hierarchal interaction [50–52]. The most recent attempts to develop holistic and integrated general theories of fundamental importance can be found in theories of self-organization that emphasize the spontaneous emergence of complex order from local interactions [53,54].

Thus cosmic evolutionary theory may be usefully applied in order to integrate an analysis capable of understanding the futures horizon of technological singularity where we predict an emergent order from the spontaneous local interactions of scientific networks internal to general society [55]. In this evolutionary philosophy human beings (and life and mind in general) can be meaningfully situated within the totality of cosmic processes of a multi-level hierarchal interaction, as opposed to being de-centered by multiple perspectival shifts (i.e., Copernicus, Darwin, Freud) internal to reductionist and fragmented science [56]. For example, in the cosmic evolutionary worldview the astrophysical singularity origin of the universe which gave rise to matter-energy and spacetime is not only an event that can be framed and resolved by quantum cosmology [57], but also an event that can and may need to be connected historically through dynamic processes of change that are giving rise to an emergent global civilization in the 21st Century [6] (Figure 2).

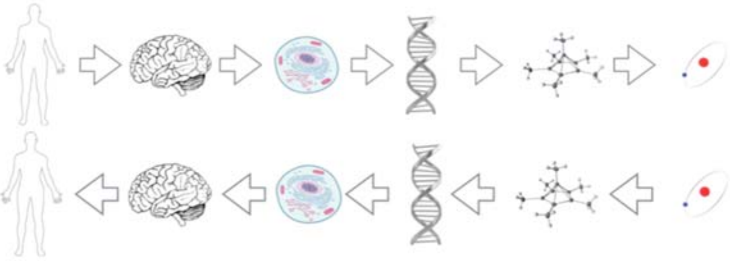


Figure 1. Reductionist versus holistic approaches. In the reductionist approach phenomena are separated and isolated into their component parts in order to understand the mechanisms unique to that level of reality. In contrast, in the holistic approach phenomena are connected and linked into networks or groups in order to understand the qualitative properties that emerge at higher level orders of reality. From the reductionist approach higher level phenomena are reducible to lower level phenomena, and from the holistic approach higher level phenomena can only be understood on the terms of their own emergent properties.

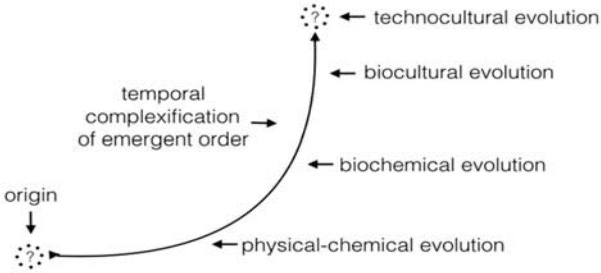


Figure 2. Big historical view of cosmic evolution. Throughout the history of universal evolution (change over time) phenomenon have undergone qualitative transitions enabling the production of new possibility and new order that operate under different basic rules or logics of change. In this schema two basic mysteries connect the beginning (alpha) and end (omega) of cosmic change with the origin of all known matter-energy and spacetime; and with the fate of the most complex known process which gives the signal of a potential future qualitative transition towards new possibility and new order.

These processes of change can be conceptualized in terms of a chain of rising complexity that generates qualitatively novel regimes of emergent order [58]. This is perhaps a more productive way to understand totality as opposed to classical conceptualization between different epistemological fields of study; or even between different ontologies of nature-culture, materialism-idealism. These regimes of emergent order can in turn be studied structurally from the simple origin of fundamental sub-atomic particles mediated by the forces of nature to the modern world of complex cognitive and social interactions mediated by the forces of ideation [59]. The logical next step would be to understand the nature of the rise of complexity and its ontological ordering consequences in relation to contemporary civilizational dynamics.

In order to approach this problem we must first clearly define what we mean by the term complexity. Complexity refers to phenomena that are fundamentally interconnected, enmeshed, and/or entangled in organizational networks of cause and effect [60]. The level of complex phenomena can be measured systematically by identifying the nature of the distinctions (meaningful differences) and connections (linked nodes) that define the organization from the lowest levels of physical order (strings, quarks, neutrinos, etc.) to the highest levels of social ideational order (languages, cities, cultures, etc.) [61]. Here we can say that a distinction introduces a division into being, whereas a connection introduces a unity into being. Thus, the analytical use of complexity in the cosmic evolutionary worldview allows for the situation of a clear and unified narrative frame of relational phenomena on all scales or levels from the sub-atomic [62] to the social [63].

Consequently, general theorists can use this frame to identify an increase or decrease in complexity when there is a change in the nature of the differentiated distinctions (divisions) and integrated connections (unities) that produce and characterize the qualities and intensities of the systemic organization [64]. For example, astronomers may identify that the universe began a complexification process when galaxies emerged because we can observe phenomena capable of higher connection (stars) through higher differentiation (chemical elements) than possible in the primordial universe. This process in turn birthed new qualities of spacetime curvature, heat radiation, photon emission; and also new intensities of nuclear fusion producing heavier substances [65]. The power of cosmic evolution to produce emergent qualities and intensities through connections (unities) and distinctions (divisions) is perhaps its most paradigmatic aspect [66].

The general trend in this complexity increase reveals a distinct arrow of time (from past to future via the present). This arrow is revealed because the mechanisms utilized by emergent phenomena to order itself are irreversible (future directed, from past to present) [67]. We may say that the temporal asymmetry of emergent phenomena is one of if not the most important distinctions between cosmic evolutionary philosophy and reductionist physical philosophy which presupposes the existence of temporally reversible eternal laws [68]. Consequently, the arrow of time can be described as an irreversible work driven by energy flows of particular dense and ordered configurations of material phenomena. These complex ordering phenomena are capable of overcoming the probabilistic statistical tendency of the universe towards greater levels of disorder or randomness [69].

In other words, because there are more ways for phenomena to be disordered than ordered (probabilistically), in the absence of local work energy directed towards more ordered states, phenomena will tend towards maximal disorder [70]. Thus when we look out at the history of an interconnected hierarchical universe and see galaxies, life, and mind in ordered configurations (instead of random organizations), we are looking at emergent patterns that evolved due to inherent internal tendencies to maintain order against a background that tends to void (Figure 3). This is true whether order manifests itself via the attractive force of physical gravitation which builds the reality of solar systems, the attractive force of biological fitness which builds the reality of ecosystems, or the attractive force of ideational desire which builds the reality of historical symbolic-systems. When contemplating the reality of physical solar systems, biological ecosystems, or historical symbolic systems we can thus think of all as part of the same cosmic evolutionary ordering force.

Thus the central remarkable aspect of the rise of complexity throughout cosmic evolution is that far-from-equilibrium organizational manifestations (i.e., changing order) tend to preserve themselves against the tendency to thermal-energetic equilibrium (i.e., maximal disorder) [71]. These organizational manifestations appear to achieve this feat through the evolution of increasingly sophisticated mechanisms of processing information which enable higher possibility spaces for actual development. In this sense we could conceptualize the arrow of time as concentrating itself multi-locally throughout the universe as a totality towards the virtual future with the potential aim of 'maximal connection' (highest unity) via the 'clearest distinction' (deepest division) given the virtual state space of the actual [72]. In our current philosophical notions of totality we have no idea how to consequentially integrate this reality into our conceptions of fundamental theory although it directly impinges on the futures reality of global technological singularity.

In our analysis we may say that the most fundamental aspects of these complex ordering phenomenon is that their identity is incomplete and open as opposed to complete and closed. Thus complex ordering phenomenon are themselves nothing but radical processes of becoming different (distinctions, divisions) in relation to basins of attraction (connections, unity) that emerge virtually as a consequence of actual relations [73] (Figure 4). This means that complex ordering phenomenon possess latent virtual potentiality which is invisible when observing the phenomenon at any particular moment, but which is still a part of the actual in the sense that actual phenomenon will tend towards virtual attractors that characterize its state space. In relation to the idea of a unified cosmic evolutionary ordering force this can occur in the physical order when a nebula becomes a star, or in the biological order when a seed becomes a tree, or in the symbolic order when a self-consciousness becomes a philosopher, or on the historical level when symbolic tribes become a global civilization. In all situations we do not have complete and closed identities reversibly existing from eternity but incomplete and open identities irreversibly existing in an asymmetrical temporality. In other words we have identities that are nothing but a process of becoming.

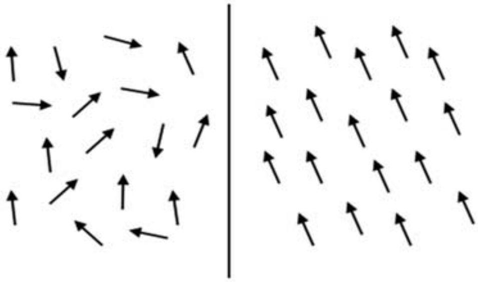


Figure 3. Disorder and order of phenomenal telos. For a disordered phenomenon (left) there is no discernible regularity of motion to be found in the constituent elements of the system as a whole. Consequently, disordered phenomena do not produce far-from-equilibrium emergent order. For an ordered phenomena (right) there is a discernible regularity of motion (a type of 'pattern') that operates on particular rules or logics which enable the phenomenon to dynamically maintain itself over the course of time (e.g., biological information of DNA, linguistic information of language are good examples of ordered phenomenon that maintain themselves with particular rules or logics). Consequently, ordered phenomenon are capable of producing emergent phenomena that are in turn capable of maintaining themselves further and further from disorder (i.e., 'higher orders' like biology from chemistry, or like symbolic from the biological).

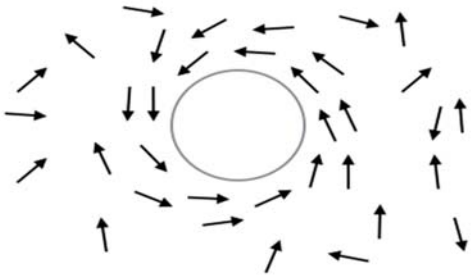


Figure 4. Becoming in relation to basins of attraction. Ordered phenomenon tend to have a collective direction of motion in relation to basins of attraction where all or most constituent parts orient informational function towards a goal-object that could not be achieved by any one part independently. For example, on the level of physical order a solar system only forms when planetary bodies orient motion around a collective center of gravity; on the level of biological order a cell can only form when chemical components orient motion around a collective central chemical code; on the level of symbolic order a society can only form when conscious components orient motion around collective central linguistic code.

In the human symbolic realm this potential aim appears existentially on the ideational horizon in the form of universal ideality (i.e., dreams, fictions) [74]. These dreams or fictions can either represent attractive or repulsive virtuality that exist phenomenologically but do not exist in the actual as observable. In other words the reality of this virtuality is only as a potentiality of consciousness in the sense of a state space with capacities that depend on the finite actual for embodied and embedded motion. For example, for both human individuals and collectives a dream or fiction structures the becoming of process, which can either relate to the individuation of a self-consciousness or the collectivization of a symbolic system. Thus, dreams and fictions in the human world are not merely epiphenomenal but fundamental to the emergent development of our realm. In their most extreme and powerful metaphysical expressions dreams and fictions may be described as future-oriented desires driving repetitive distinctions (divisions) attempting to form higher connections (unities). On the terms of individual self-consciousness this may be conceived of as something like the 'overman' (higher self); and on the terms of collective social systems this may be conceived of as something like 'utopia' in secular context (higher civilization), and 'heaven' in transcendental context (higher being). In all cases we are dealing with the nature of the highest or deepest values, purposes, and meanings as emergent basins of attraction which structure phenomenal historical becoming.

This cosmic evolutionary philosophy as foundational worldview could challenge our dominant cosmological conceptions grounded in thermodynamics of an imminent tendency to universal disorder (Figure 5). In this philosophy the immanence of higher orders may be currently absent, existing only as an invisible potentiality internal to and dependent on the actual intensities and qualities of a future qualitative phase transition (Figure 6). Furthermore, these actual intensities and qualities of a future qualitative phase transition could be driven by psychosocial processes that are equally fundamental to lower order reductionist physical models of reality that cannot approach the irreversible temporality of complex ordering phenomenon. Or it could be that reductionist physical models are necessary conceptual structures for the actualization of the next qualitative phase transition. The clear location of this new complex order should be concretely related to basins of attraction of the differentiated distinctions related to human beings (divisions) and the way in which their psychosocial intensities and qualities are coordinated and constrained in integrated connections (unities). In the global technological singularity literature the future divisions are often represented as artificial general intelligences and the future unity is often represented in terms of distributed digital networks [75,76].

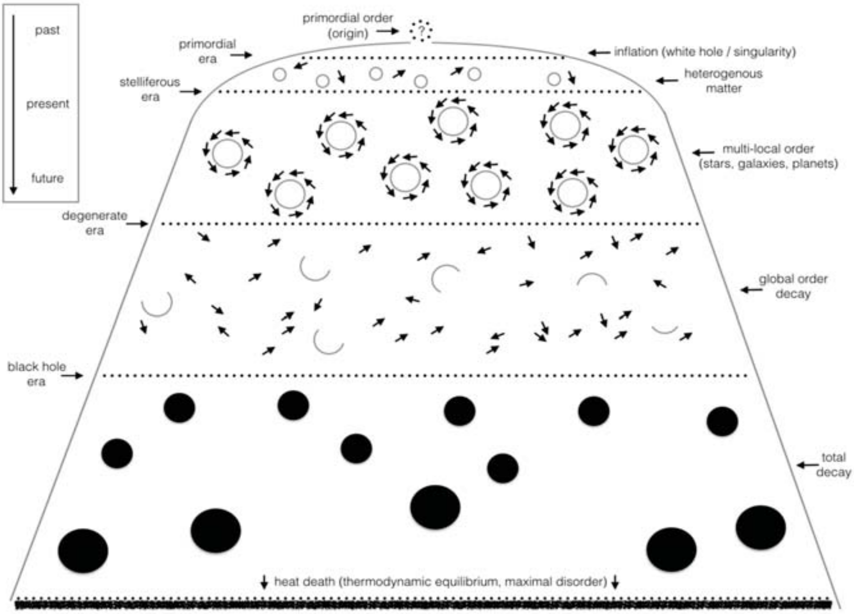


Figure 5. Thermodynamics view of the cosmos, primordial order to final disorder. The thermodynamics view of the cosmos gives the picture of a universe with particular low-entropy, highly ordered or supersymmetrical initial state of being (non-random motion). This initial state drifts towards higher-entropy, global disorder (random motion) over time via symmetry breaking events (divisions) and feedback loops (unities) which generates a motion that we understand as an arrow of time. Consequently, in the context of the universe as a whole (considering the whole of space and the whole of time) the most common state space for matter is general disorganization (thermal equilibrium) due to low material interaction rates, which suggests that the currently observed state of the universe is ultimately unstable. The multi-local material order that does self-organize into persistent temporal form (galaxies, stars, life, mind) occurs due to gravitational attraction acting on heterogeneous distributions of organization which enables higher material interaction rates. In our current understanding of the universe there is no complete theory that explains the fundamental consequence of the emergence of such multi-local order, and reductionist perspectives tend to regard such phenomena as epiphenomenal. In other words, reductionist perspectives identify a fundamental objectivity (unity) framed a priori by a subjectivity (division), but cannot think a framed a priori subjectivity (division) that constitutes an emergent fundamental objectivity (unity).

From the level of phenomenology this places importance on experiential horizons of reality that are more fundamental to being than any particular forms of scientific knowledge (i.e., thermodynamic theory) [77] (Husserl 1970). These horizons function as new levels of appearance for interaction and as new platforms for potential complexification to be ordered by the future-oriented cognitive agents under the particular attractors and constraints unique to that level of becoming [78]. Thus the cosmic evolutionary worldview suggests that our universe possesses a denser concentration of relation and a more lively cognitive destiny filled with more adventure and mystery than is often conceived in dominant conceptual paradigms [79]. In this way the thermodynamic worldview positing an inevitable tendency to maximal disorder or non-being [80], may be missing a crucial part of the picture: the nature and future immanent potentiality of ideality inherent to the symbolic order of being [77,81]. If we include this potentiality we must also consider the possibility that there exists a to-be-actualized singularity of ideational order that cannot be thought in the thermodynamic conception of the cosmos.

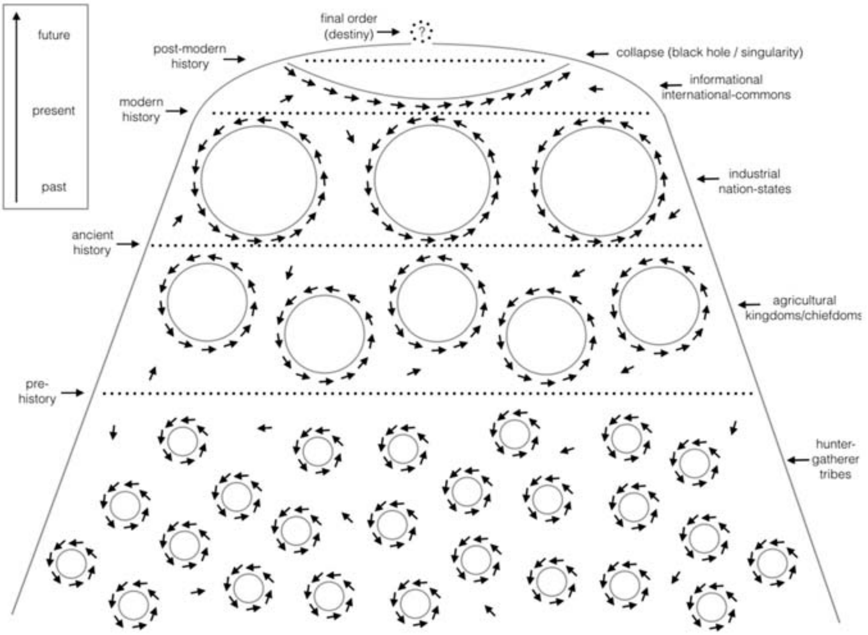


Figure 6. Teleodynamic view of the (local) cosmos, primordial disorder to final order. The above representation attempts to capture the cosmic evolutionary worldview that is characterized by far-from-equilibrium or non-equilibrium systems that operate on self-organizing principles dynamically balanced between chaos and order. In the teleodynamic conception we get an image of the world that presents us with an immanent 'immortal heat' where highly ordered far-from equilibrium systems curve their being to a state of supersymmetrical unity (a cosmic-transcendental monism). Such a state would likely annihilate the dualistic distinctions between subject-object, concept-world, observer-observed, material-ideal without resorting to a pre-linguistic 'biophysical grounding' that ignores the emergence and consequences of conceptual distinctions (i.e., 'distinction-division dynamics'). In this representation the totality of process is conceived of as starting with the emergence of a field composed of ideationally constituted social unites (bands/tribes) whose ground is self-consciousness developing in language. Throughout the historical process bands/tribes become progressively 'synthesized' into higher level social unities which has the effect of reducing the number of different unified groups (i.e., fewer unities) but increasing the spatial scale of the unified groups (i.e., the difference between Europe pre-and-post Roman Empire, or the Asia pre-and-post Chinese Empire, etc.). In this progressive trend to unification the level of individuation also progressively increases meaning that there are emergent degrees of freedom for the particular elements of the higher level social unities. This paradox between higher social unity and higher individuation continues to the present day where we see the dominance of a 'multiplicity of ideals' which are nonetheless all expressing ideality within one universal technological medium. The combination of these two trends make it difficult for philosophy to make sense of totality. In this view in order to approach totality we must include the radical divisions characteristic of individuation into the higher unity of totality, thus creating a unity inclusive of division.

3. Dialectical Approach to Technological Singularity

In order to continue this analysis we here shift focus from the object-oriented cosmic evolutionary philosophy itself towards understanding how we can utilize this philosophy dialectically as it relates to an internal subjective approach to global technological singularity. The objectivity of human motion mediated by universal ideation is not fully understood in either the classical mechanistic worldview (scientific materialism) or the classical transcendental worldview (philosophical idealism). This is because scientific materialism reduces everything to physical motion governed by eternal laws, and philosophical idealism integrates everything into the eternal concept which transcends physical reality. Thus both approaches fail to capture the dynamical intersubjective becoming of ideality and its historical effects and consequences in relation to totality. In order to approach ideational motion embodied and embedded intersubjectively in a material world we must understand the real consequences of an emergent subject-object division (or observer-observed, mind-matter, concept-world) as itself a dynamical part of totality (i.e., freedom is part of totality).

When we start with a totality constituted by subject-object division (as opposed to starting with a subject-less objectivity, i.e., big bang; sub-atomic realm, etc.) we are starting with the asymmetrical becoming of our own phenomenal existence which appears to freely aim for a higher unity (or 'oneness') on the individual level of self-consciousness and on the collective level of social systems ('self-actualization'). The subject-object relation is of fundamental importance when reflecting on a global technological singularity as an objective real event. This is because we must understand how objectivity itself depends on the activity of open and incomplete observers manifest as intersubjective networks striving for self-actualized unity within multiplicity. To approach this phenomenon we will rely on a philosophy of materially grounded idealist dialectics, and we will rely on the science of psychological self-analysis whose focus is the objective content of ideal becoming. In other words, what we get from a synthesis of scientific materialism and philosophical idealism is a dynamical intersubjective objectivity with a locus of causality internal to conscious ideality as fundamental reality [82].

The mediation of an ideal intersubjective objectivity gives us a potential human-centric perspective on approaching technological singularity theory where human observers are the active psychosocial drivers of process. To be specific this dialectic operates on general negations of the world as given (divisions) coupled with affirmations of an alternative idealized world (unities). In the present the conceptual motion of negation and affirmation occur within the unified medium of the internet which enables both a higher level of individuation (division) and a more totalizing universality (unity) than ever before in human history [57,78]. Here ideal negations and affirmations are phenomenologically tested in relation to the actuality of their enacted transformative consequences. In other words, a particular individuated division repetitively attempts to introduce and maintain the actuality of a simple unity in being against a complex background. In this general motion of ideality transformations retroactively change the subject-object relation, and thus totality itself, by creating and mobilizing psychosocial processes towards new ideal realities [83]. In order to approach this divided field, a general motion can potentially be formulated under a schema of triadic logic [44,84] from (1) theoretical abstractions (imaginary) to (2) enacted transformations (symbolic) to (3) concrete actualization (real) (Figures 7–10).

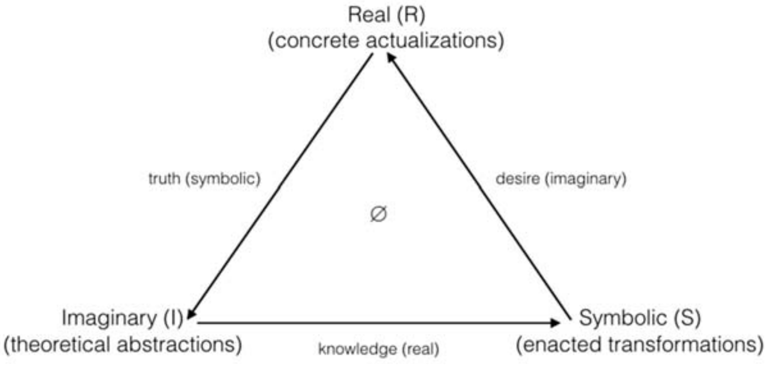


Figure 7. Structural transformations on the transcendental horizon: triad. This structure represents an attempt to formalize the geometry of the mental space which becomes entangled with the world. From this entanglement we see the motion of the dynamical geometry of the mental space as concretizing new reals from knowledge by processing imaginary desire through symbolic truth acts. First, the subject that intervenes/transforms emerges from the concrete real in relation to an imaginary field or screen. This field or screen becomes the invisible location of the subject's thought on its perceptual horizon of becoming and reveals to the subject an object of desire which motivates it to intervene/transform the real towards a more ideal order. The subject achieves such interventions/transformations through introducing the symbolic to the real in the form of speech, writing, poetry, song, equations, jokes, and various other formalisms. From introducing the symbolic into the real the subject becomes fundamentally entangled, correlated, and/or caught up in a web of interactions with the real and thus measures its symbolic on the terms of which the real returns its message as truth (intervention/transformation). In these processes all three elements of the imaginary-symbolic-real are simultaneously co-present: the real of knowledge is present when the imaginary theoretical abstractions collapse into enacted symbolic transformations; the imaginary of desire is present when enacted symbolic transformations attempt a real concrete actualization; and the symbolic truth is there when the real of concrete actualizations informs theoretical abstractions of the imaginary. In this scheme 'knowledge' is 'real' because knowledge is the concrete actualization of the subject, 'desire' is 'imaginary' because desire represents an absent actualization, and 'truth' is 'symbolic' because truth is what returns to the subject from the real (an 'answer' from the real that 'emerges' due to our own interventions/transformations). In order to interpret and utilize this structural triad it is crucial that one conceives of it as always-already ('eternally') in motion and constantly re-setting its coordinates in a network of such structures which compose the geometric space of the four-dimensional transcendental horizon. The central Ø denotes that 'beyond' or 'behind' the transcendental horizon there is nothing other than the void of our subjectivity which we project (imaginary) and enact (symbolic) into the real. In this way such an analysis attempts to think a dynamical real virtual thing-in-itself (between order-chaos, dreams-obstacles, goals-challenges, presences-absences etc.) that is both open and incomplete making the 'thing' dependent on subjectivity for its own identity and becoming due to a division that is fundamental for the appearance of subjectivity as such.

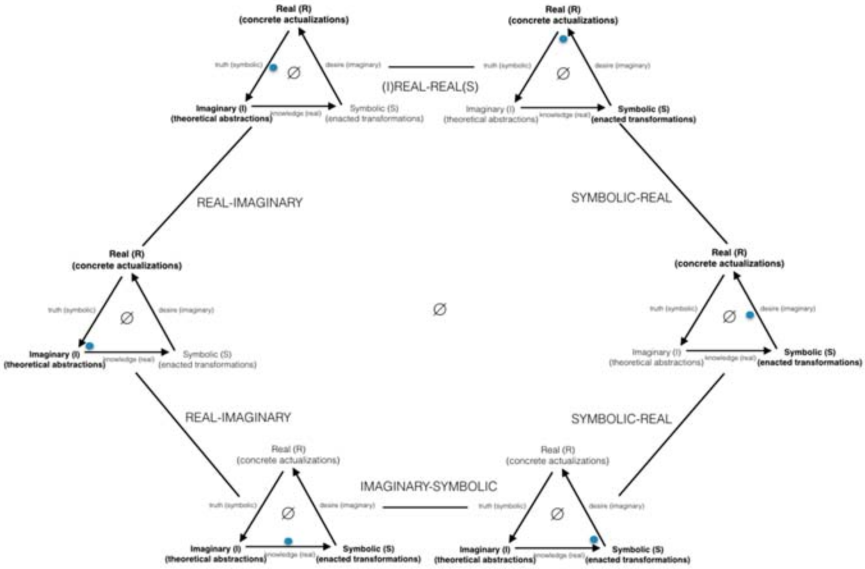


Figure 8. Structural transformations on the transcendental horizon: cyclohedron. The geometry of the triad is a classical structure of logic used to study the nature of the human mind (e.g., for psychoanalysis: [85]; and philosophy: [86]; see also: [87]). However, in order to capture a larger view of totality in triadic transformation we must also represent the dynamical motion with a higher dimensional cyclohedron representing the circular-orbital temporality of desire (blue dot, objet a) as it passes from imaginary-symbolic, symbolic-real, and real-imaginary. Cyclohedrons complexify triads in geometrical analyses used to study knot invariants ([88,89]). Here knot invariants are desire deadlocks (points of impossibility) fundamental to understanding the structure of pure desire. In this frame 'knots' can be located on the level of the symbolic transformation which generates a perceived subjective separation/connection (in space) and distance/closeness (in time) from the 'real event' as the 'desired state' (ideal-real). In other words, when symbolic transformations 'miss' on returning from the real there is a subjective emptiness (filled only with the real of imaginary potentiality), and when symbolic transformations 'hit' on return from the real there is a subjective fullness (where the imaginary intersects with the real actuality of desire).

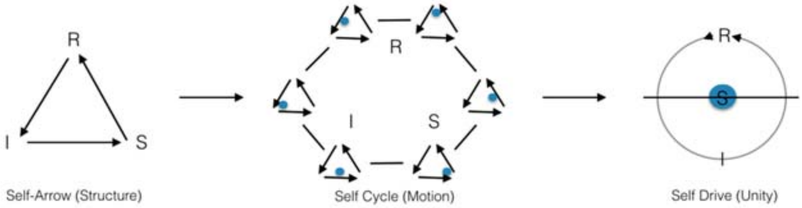


Figure 9. Structural transformations on the transcendental horizon: not-One. In the higher orders of Self we find a Unity in the not-One where the Real of openness and incompletion is fundamental to the Self's internal Unity. The inner discovery of this Unity shifts the subject from a mode of pure desire to pure drive. In the mode of pure drive the gap/difference the Imaginary detects in the Real produces a Symbolic rotary motion of pure enjoyment (ideal-real) where desire is 'eternal'.

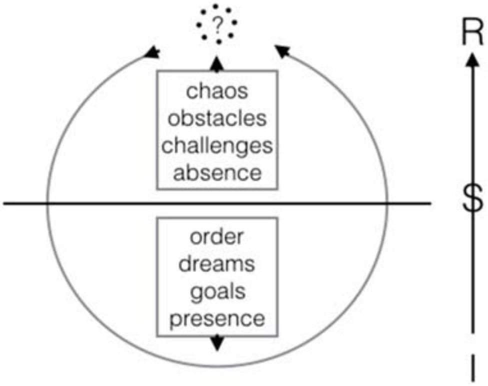


Figure 10. Totality as not-One, dynamical structure of the transcendental horizon. The totality or 'thing-in-itself' is not only 'real virtuality' but in-itself a dynamical becoming (making it 'not-One'). This not-One is thus irreducibly dual (subject-object, concept-world, human-nature, mind-matter, etc.) and split between ontology/being that resists and epistemology/knowledge that insists. On the ontological side of the 'impossible thing' which resists our grasp we may place the domain of chaos, obstacles, challenges, and absences; and on the side of the 'transforming consciousness' we can place the insistent grasping domain of order, dreams, goals, and presences. The subject caught up in the triad of the imaginary-symbolic-real of this process (self arrow) is what dynamically traverses both sides of the dividing line between order-chaos, dreams-obstacles, goals-challenges, presence-absence in an asymmetrical future-directed motion. In other words the subject is what invades being with epistemological insistences ('projects'), and experiences in return ontological resistances. The subject attempts to bring order to chaos, achieve dreams that overcome obstacles, reach goals that nihilate challenges, and to make present which was once absent through its symbolic transformations. In analysis of this open and incomplete 'thing-in-itself' can we possibly think the actualization of the thing-in-itself as One or is totality nothing but reducible to this impossibility as not-One? What is in the end the status of this gap/difference internal to the thing-in-itself?

Thus, in focusing on an ideational motion structured by triadic logic operating internal to a field of multiple observers totality itself becomes open, incomplete and dependent on intersubjective ideality for its actuality. Thus in this frame totality itself is nothing but the ideational space of observers where epistemological knowledge practices become entangled with fundamental ontology. Here knowledgeable observers become effective agents that are directly involved in transforming the physical world via conceptual mediation. From such a program we are invited to engage with a spatially curved epistemology due to the inclusion of desiring subjects that develop strategies for coping with internal symbolic deadlocks produced as a consequence of division from unity (or 'points of impossibility'). We are also invited to engage with a temporal dynamic ontology due to the inclusion of the ontic effects of subjects own projected interventions and retroactive reflections aiming for unity. This intersubjective motion is structured by a multiplicity of ideational networks capable of negating-affirming (transforming) the distance between the actuality of friction (zero-sum) and the ideality of synergy (positive-sum).

In this frame totality becomes a radically asymmetrical temporality of ones dualistically imbalanced between being and absence [90], order and chaos [91], dream and obstacle [92], goal and challenge [93]. Thus, in terms of subject-object (observer-observed, mind-mater, concept-world) division we have clear separation and distance from any understanding of totality as eternal unity (physical/material or conceptual/ideational). On the side of the subject, we find our being immersed in nature, we find an order that frames the world, we find a dream that orients action, we find a goal to transform the given. On the side of the object we find an absence in being, we confront an unknown chaos, we find that there are obstacles on our path, we find that movement is a challenge. Such separation and distance represents an eternal division that disturbs the possibility of eternal unity but allows for the asymmetric temporality of free representations in historical becoming. Here totality is not closed and complete but in need of a reconciliation which involves the active participation of every observer as a background independent field of ones (individuals). Such reconciliation can be framed experientially as a desire to enter a flow state or drive that would actualize a qualitative phase transition towards a higher order state of unified being.

When thinking totality as an eternal division to be reconciled in these terms we may posit that a future actual singularity is immanent in the potentiality of our action even though we have no idea what will manifest on this pathway. Here consider the general epistemological structure of scientific networks as forces that aim to transform being and retroactively change the way in which we think about subject-object division as totality. The most obvious historical examples of this retroactive change include the industrial technology revolution which changed our conceptions of the energetic cosmos (i.e., thermodynamics, etc.), and also the computer technology revolution which changed our conceptions of the informational cosmos (i.e., cybernetics, etc.). These examples are so general that when we are thinking about the desiring subjects who drive the future of robotics, nanotechnology, genetics, quantum computing, and so forth, it is hard not to conclude that the logical immanence of their own sets of ideal conceptualization schemes will retroactively change our fundamental conceptualization of reality on the level of universal history [94]. Indeed, it could be that fundamental reality or totality depends on the intervention of conceptual observers and their technic apparatuses in order to be constituted and thus realized.

Of course in this process not all modes of individuation are equal because not all individuations lead to effects of universal historical significance. Consider an extreme example of individuation with consequences for universal history with the most (in)famous technological singularity theorist, Ray Kurzweil, and his sets of ideal claims about future immortality [20,21,95,96] (which emerge against the background of a fear of death (i.e., separation and distance from eternal unity)). In a dialectical analysis of the higher order ideational space the point would be to understand the effects and consequences of this ideality in terms of their abstract powers to transform being for all other individuated observers. These abstract powers include the force of a paradigm that can directly or indirectly create and mobilize psychosocial environments of potentially extreme relevance to universal history. For example, in the case of Kurzweil's set of conjectures we have psychosocial systems as unities focused on creating the next generation of artificial intelligence. These psychosocial environments have gained a symbolic autonomy in chains of communicative events which are unlikely to be prevented from exhausting their potentiality. The nature of subject-object division will be forever transformed as a consequence but what dynamical state of phenomenal being will they actualize?

Consequently, all processes of individuation are crucial for the reconciliation of universal history on the highest orders. This is for the simple fact that all subjects are an effect of the eternal division and all subjects are ultimately responsible for producing an internal unity capable of withstanding this asymmetrical imbalance. In thinking about the technological singularity with such dialectical tools we aim to reframe the position of humanity in the philosophy of science and the sciences of humanity as a whole. This ultimately requires us to understand the motion of general ideation as it will be effected by scientific ideational reduction and fragmentation as they are currently manifesting and evolving spatially in curved virtual fields [97]. These virtual fields include ideational forms as diverse as M-theory in particle physics, artificial intelligence in cognitive science, and genetic engineering in biology. All such fields could be technically modelled from the point of view of a temporally asymmetrical totality in a state of subject-object division. The collective nature of this state space suggests that the actuality of human civilization must include within it an invisible potential tendency to transcendence of present being that can be discursively mediated in relation to a central desire for higher order unity.

In order to demonstrate the functional or pragmatic utility of such a seemingly complex notion of totality let us now consider an example of how this approach can help us understand the higher orders of scientific ideation operating under paradigms of fragmented reduction. The example I will choose in this analysis of totality is the example of the theoretical particle physics community. The reason to consider this community as a particular example with general relevance to global technological singularity theory is that this community is the best example of fragmented ideational reduction applied towards a unified theory of totality. In contemporary particle physics theory ideal abstractions conjecture a theory of quantum gravity capable of understanding physical singularities (black holes, big bang). In order to realize this the theoretical particle physics community holds a particular set of ideality about the nature of the world (i.e., standard model, M-theory). These abstractions aim to understand the hypothetical dimensions that may represent the fundamental constituents of physical matter at the smallest scale of reality.

Now when we are thinking about the particle physics community as a psychosocial force produced as an effect of subject-object division aiming for unity we have to think about how this asymmetrical imbalance generates ideational conflict and tension. This conflict and tension can be situated around the aforementioned dualistic eternal couples of totality: being-absence, order-chaos, dream-obstacle, goal-challenge. Here we see that the particle physics community emerges as an epistemological desire that will retroactively change ontology (i.e., standard model is incomplete logic, M-theory is open speculative conjecture, etc.). Thus, in the higher order theory where we include the observational multiplicity as fundamental we could posit that abstractions in the form of the standard model and M-theory enter into a negative relation with given being. This occurs in so far as the projected ideality of M-theory posits a future reconciled or completed state (imaginary) within which totality will be transformed towards a comprehensive notion of the fundamental constituents of physical matter (symbolic) which would allow us to understand the origins and fate of all matter (real).

However, on the first order of analysis there has been little progress in understanding quantum gravity in terms of concrete predictions. This is due to the inability to test the real of projected ideality that would help us understand the origins and fate of all matter. The result of such obstacles leads to fundamental theories disconnected from the classical reality of scientific materialism and much more connected to the classical reality of transcendental idealism. Thus, in order to avoid both eternal structures it could be that a higher order perspective emphasizing the eternity of a subject-object division attracted towards an emergent unity constituted by intersubjectivity could be what is missing from our consideration of fundamental problems related to quantum gravity. In the higher order frame the solution to quantum gravity must contend with the real becoming of observers constituting the transcendental horizon. In other words a higher order theory of quantum gravity would not only have to explain the nature of the big bang and black hole singularities, but also the nature of conscious observers in the potential actuality of a technological singularity (Figure 11).

In this context the particular example of analyzing the particle physics community and their first order search for unified totality becomes extracted to the higher order and applied to general psychosocial systems. This seemingly strange move in relation to understanding unified totality has fundamental implications for reductionist theories that attempt to explain everything. For example, a common feature of modern scientific discourse is the claim that a grand unified theory is within our grasp. However, when we consider this proposition in our present context, is it even possible for us to imagine what a grand unified theory would actually mean for the subject-object division as totality? What would it mean if communities of physicists developed a grand unified theory of quantum gravity as a reductionist master theory of the universe? In other words, would such a theoretical development (imaginary) have any practical transformative consequences (symbolic) for the emergentist subject-object division as totality (real)? Would it mean that something fundamental would change for the nature of temporally asymmetrical and irreducibly perspectival discursive reality?

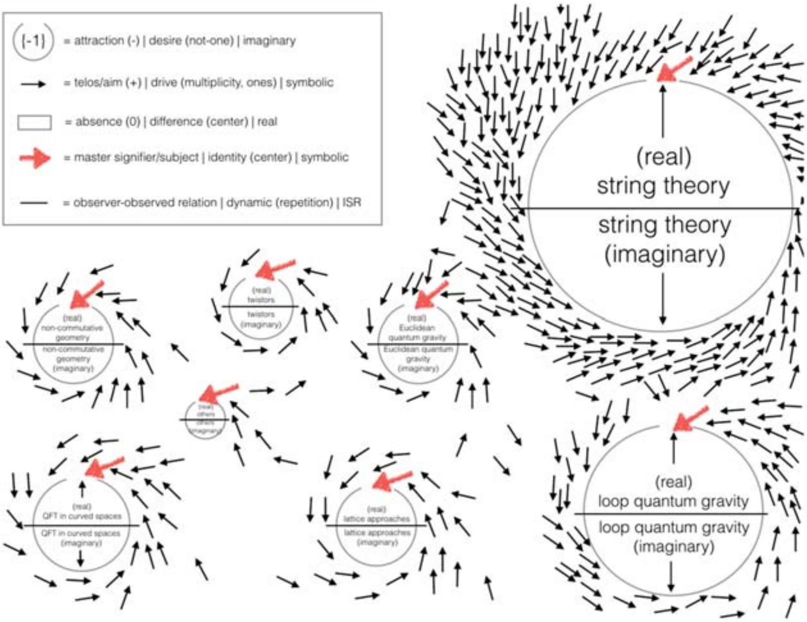


Figure 11. Psychosocial virtual field of physics communities. The above figure attempts to capture the general psychosocial virtual field of quantum gravity in relationship to the main abstract theoretical groups (imaginary) which aim to transform the concept-world relation through enacted transformations (symbolic) by concretely actualizing the 'true' understanding of quantum gravity (real). The red arrows represent a 'master signifier' or 'main subject' which could be dominant concept(s) or person(s) in the respective fields (i.e., 'strings' and 'Ed Witten' or 'loops' and 'Lee Smolin') which orient the symbolic weight and aim (topography of black arrows) of the ideational field in relation to the absent real. The closer the network of arrows/signifiers to the real the closer they are to the in-itself of the social-historical meaning of the semiosphere. For example, black arrows (symbolic transformations) close to the site of the real may be fellow professors or graduate students with the trailing arrows representing lower levels of relation to the site of the real, i.e., colleagues, other scientists, educated public, undergraduate students, etc. In these relations the bar of the circles represents the dynamic border of the 'thing' (from order/dream/goal to chaos/obstacle/challenge and back) that can be symbolically conceptualized in light of the linguistic formalisms S | s denoting the signifier's freedom and autonomy over and above the signified [98]. This is best demonstrated by physics communities whose signifying chain tends towards the in-itself of internal-imaginary consistency before it is validated by the return from the real. These represented social systems can be divided down the 'three main roads' of quantum gravity, i.e., (1) approaches that start with quantum mechanics (e.g., string theory, QFT in curved spaces), (2) approaches that start with general relativity (e.g., loop quantum gravity, twistors), and (3) approaches that start from novel presuppositions that rethink the foundation of both quantum mechanics and general relativity (e.g., 'others') [99]. The specific sizes of the psychosocial gravitational fields in this representation correspond to data collected by Carlo Rovelli at the International Congress on General Relativity and Gravitation on the count of articles published in each respective field of quantum gravity [100]. Of course what this representation does not capture (requiring more sophisticated modelling) is the dynamic motion, the interaction, and the change over time of these communities. Such a model could potentially approach questions about why many different socially constituted physics tribes seem to be able to produce correct solutions to the origin and fate of matter with different fundamental constituents of quantum gravity [101] and thus also potentially help produce a higher order relational understanding of reductionist attempts at complete-closed grand unified theory modelling.

In proper philosophical terms a grand unified theory can be conceived as an eternal unity often represented on the level of understanding as absolute knowledge [102]. However, this unity would logically nihilate the primacy of the subject-object division whose broken symmetry is causative of ideationally constituted intersubjective becoming. Consequently, whenever we are approached with first order notions of a grand unified theory we should never forget that these abstract conjectures central weakness is their inability to deal with the (still moving) higher order psychosocial forces where the subject-object division is left without total reconciliation. Thus, when it comes to theories of quantum gravity it may be necessary to situate those dealing with subject-object division as the synthetic 'third path' in contrast to the first path grounded in reconciling general relativity with quantum mechanics (e.g., string theory); or the second path grounded in reconciling quantum mechanics with general relativity (e.g., loop quantum gravity). In the third path what becomes crucial and indispensable is the asymmetrical nature of time and the meaning of a universe in which we are active participants [103].

4. Consciousness and Universal History

We have attempted to build a cosmic evolutionary philosophy and situate within this philosophy the fundamental dynamics of ideational motion on the horizon of universal process. Now we will attempt to situate ideational motion within a higher order theory of consciousness that can approach totality. In this theory of consciousness, we place less emphasis on the physical instantiation of consciousness within a materialist foundation and instead place more emphasis on historically-engaged phenomenal understanding as it relates to a fundamental truth of unified reality. In other words, this analysis is less concerned with whether consciousness is produced by neuronal activity, or by the quantum level of being, or by some other unknown physical mechanism; and is more concerned with the phenomenal activity of psychosocial forces as they relate to the historical search for the truth of reality.

In building this theory of consciousness, our analysis will forward a different perspective then most reflections since it will offer an emergentist mental theory, as opposed to a scientific reductionist theory or philosophical transcendentalist theory, seeking to understand totality in terms of its relevance to the meaning of human existence. This approach to totality is different than most contemporary theories because, instead of explaining totality in terms of the mechanics of sub-atomic reality or the eternal absolute, we are explaining totality in terms of general ideational motion engaged in history (like in the analysis of the physics community). Thus, we are interested in a totality capable of helping us understand how frames of reference and their conceptual transformations will be generally effected by scientific epistemology (artificial intelligence, genetic engineering, quantum computation, etc.). In this way, we seek to emphasize the hard work of an emergent unity or integration via collective historical processes of human individuation. Consequently, this theory of consciousness aims to elucidate a central dynamical narrative and value structure of being that is both grounded in cosmic evolution as a universal process (as emphasized in Part 2) and future-oriented towards a meaningful synthetical higher order level of ideational order (as emphasized in Part 3).

The first step in constructing a theory of consciousness with relevance to universal history is to situate our understanding within proper historical context. In order to move in this direction, let us first consider the main metaphysical systems of thought that have structured the history of philosophy. These main metaphysical systems of thought will be broadly classified from the Western perspective as ancient, modern, and deconstructionist metaphysics. In this general classification, we can say that ancient metaphysics structured the development of civilization in its predominantly agricultural phase; modern metaphysics structured the development of civilization in its predominantly industrial phase; and deconstructionist metaphysics has structured the development of civilization in its predominantly informational phase. Thus, these metaphysics represent the logical ideational substructure of civilization at different moments in the collective becoming of self-consciousness in universal history.

In this analysis, we will pragmatically utilize the dynamical triadic structure of the imaginary-symbolic-real from Section 3 in order to situate our analysis of each major phase of Western civilizational metaphysics. Here, the imaginary as theoretical abstractions, the symbolic as enacted transformations and the real as concrete actualization in its general psychosocial manifestation are applied to different historical conceptions of unity. From all three phases, we can generalize the human mind as situated on the level of the symbolic order because self-consciousness is a narrative construct organized with a symbolic architecture capable of enacting historical transformations. However, what is considered imaginary (i.e., a theoretical abstraction) and real (i.e., a concrete actualization) will fundamentally change in all three major movements of civilization from ancient to modern to deconstructionist:

□ The ancient metaphysical structure considers as 'real' the 'eternal ideal' or God, and considers as 'imaginary' the 'physical world' or nature. Consequently, in ancient metaphysics, we get philosophies built around the ideals of a transcendent supernature that is primary in constituting the physical world and primary in relation to the human mind. Thus, ancient metaphysical systems forward the hypothesis that human beings come from an eternal ideal superspace before birth, return to an ideal superspace after death and are structured-constrained by an ideal superspace during existential (sexual-personal-creative) development.

□ The modern metaphysical structure considers as 'real' the 'physical world' or nature, and considers as 'imaginary' the temporality of the ideal. Consequently, in modern metaphysics, we get philosophies built around the natural world governed by eternal physical laws and ideas that have no 'transcendental' reality outside of their constitution in history. Thus, modern metaphysical systems forward the hypothesis that human beings come from nature before birth, return to nature after death, and are structured-constrained by the laws of physics during existential (sexual-personal-creative) development.

□ The deconstructionist metaphysical structure considers as 'real' the 'secular power' structures of society, and considers as 'imaginary' the various possible interpretations of the 'physical world'. Consequently, in deconstructionist metaphysics, we get philosophies built around the negation of social systems that seek to totalize human existence and distort our relation to the natural world. Thus, deconstructionist metaphysical systems forward the hypothesis that human beings come from social systems, return to social systems after death and are structured-constrained by social systems during existential (sexual-personal-creative) development.

This broad analysis of metaphysical totality structures aims to situate consciousness as something that is constituted by the symbolic order and constantly re-structuring its transformative state of being in relation to different notions of what is an imaginary theoretical abstraction and what is the most real concrete actualization throughout its collective development in universal history. In the ancient real, we can say that what consciousness developed in relation to was fundamentally the power of the theological and the transcendental. In the modern real, we can say that what consciousness developed in relation to was fundamentally the power of the scientific and the natural. In the deconstructionist real, we can say that what consciousness developed in relation to was fundamentally the power of the social and the self-analytic. Thus, in all systems, we get fundamentally different notions of totality as eternal unity: transcendental ideality, physical laws, or secular power.

In ancient metaphysics, totality is already closed and complete in the ideal real of supernature or God of religion; in modern metaphysics, totality is already closed and complete in the material real of natural laws of physics; and in deconstructionist metaphysics, totality is already closed and complete as a multiplicity of systems of social power (all of which represent different relativistic totalities). However, all of these metaphysical systems are unable to account for a totality where conscious reality is constituted by multiple observers becoming in asymmetrical temporal relation to ideal-real attractors independent of a transcendent superspace, physical laws or secular power. In other words, they fail to account for the general imaginary-symbolic-real triad in its own historical motion, which transcends the ancient, modern, and deconstructionist forms. Thus, for a conscious real on the level of universal history visions of totality related to a transcendental superspace, physical laws, or secular power all become a part of the same dynamical and general conscious real structuring the becoming of open and incomplete individuating observers searching for the truth of being in a unified eternal structure.

In this way, the most real, or the most concrete actualization, is an absence of 'something' that emerges because of and depending on symbolic observers' enacted transformations in history. This brings us towards a potential to formulate a theory of consciousness that can approach the real in-itself as an absence of something that emerges internal to the realm of symbolic observers. This formulation will attempt to structure a transmodern metaphysics derived from the dynamical motion of the general imaginary-symbolic-real structures (Figure 12). In a transmodern metaphysics, we aim to both synthesize historical forms of totality and approach technological singularity from an individuated perspective as an emergent unity produced as a general consequence of subject-object division. This would potentially allow us to construct a central narrative and value structure of being for consciousness. Here, narrative architectures represent a symbolical temporalization of eternity (beginning to end); and value structures represent an attempt to stabilize an emergent unity as a perfect circle capable of completing and closing in on itself.

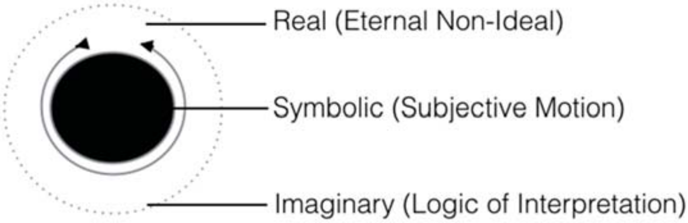

Real (Eternal Non-Ideal)
Symbolic (Subjective Motion)
Imaginary (Logic of Interpretation)

Figure 12. Transmodern totality. We represent a transmodern totality where the general motion of the human mind as symbolic emerges from the physical world as imaginary and starts to circulate around its own emergent desire for internal unity as real. The consequence of this general motion deployed in historicity produces a multiplicity of unities such as transcendental superspaces, physical laws, and secular power that stand in as paradoxical somethings where the real as a closed and completed circular is absent or impossible. In this sense all historical forms of the eternal real should ultimately be deconstructed on the level of external positing. However, all external positing of an eternal real ultimately emerge because of self-alienation as they represent the eternal real that emerges internal to the subject. In this way we find a real that is a multiplicity of unities internal to each individuated observer. Thus, the transmodern real offers the view that each subject is capable of creating its own world out of its own individuated symbolic transformations. Here the transcendental superspace, physical laws, and secular power all disappear as ultimately historical fictions on the level of specular images.

In terms of the synthesis of historical forms of totality, the transmodern real conceives of all particular cultural reals as unified conscious visions necessary for the structure of historical becoming. For example, the vision of a transcendental superspace of eternal ideality structures the objective ancient becoming of religious and philosophical intersubjectivity; the vision of physical laws structures the objective modern becoming of scientific and naturalist intersubjectivity; and the vision of secular power structures the objective deconstructionist becoming of social and activist intersubjectivity. Thus, different forms of intersubjective objectivity emerged and became necessary in different phases of civilization from the agricultural level stabilized by the intersubjective objectivity of God and Imaginary Faith; the industrial level stabilized by the intersubjective objectivity of Science and Rational Empiricism; and the informational level stabilized by the intersubjective objectivity of the Social and Critical Deconstruction. Of course, these are not the only historical forms of eternal ideality that have appeared on the transcendental horizon, but they are a few of the major and general reals that have structured intersubjective objectivity.

Thus, in terms of a transmodern metaphysics approaching the technological singularity we must not conceive of totality in terms of an eternally unified field, but instead as an eternally divided field between subject-object. As discussed this field produces individuated observers asymmetrically imbalanced in a duality universally structuring a general internal desire for unity expressing itself as a pure multiplicity of conscious visions. Here we build on the aforementioned idea that totality on the side of the subject is order-dreams-goals-presences; and totality on the side of the object is chaos-obstacles-challenges-absences. In this dualistic relation unities like a transcendental superspace, physical laws, or secular power represent intersubjective reals as 'points of impossibility' unconsciously posited by self-consciousness to resolve the far-from-equilibrium imbalance of dualistic becoming. Such reconciliation of far-from-equilibrium imbalance follows such logics as 'if we all believe in [faith, empiricism, deconstruction] we will be saved by [religion, science, society]'.

However, in terms of the transmodern real in-itself such unities as 'points of impossibility' are radically open to taking any form that an observer can maintain intersubjectively across its process of becoming as an objective reality. Furthermore, any externally imposed collectivist notion of an intersubjective objectivity will by necessity fail to approach the real of individuated becoming of a multiplicity of observers on the pathway to singularity. This is because the real as an impossibility emerges primordially in relation to divisions introduced by each subject's transformations (as opposed to preceding the subject's transformations). Thus, when we think of the real from the inside out (from the side of the subject) these points of impossibility structure a geometric curvature in a topographical state space. For example, a subject engaged in theological or philosophical transformations may conclude that the highest form of objective reality is a unified space of ideality (God); a subject engaged in scientific or naturalist transformations may conclude that the highest form of objective reality is a unified space of natural laws (Spacetime); a subject engaged in social or activist transformations may conclude that the highest form of objective reality is a unified space of secular power (State).

In all such social historical manifestations what is posited by self-consciousness is an objective real that exists before and after the subject's transformations or interventions into the real as an absolute background dependence. Thus, the historical subject tends to reify an object that it believes existed before it, and believes will exist after it (i.e., God, Spacetime, State). However, what the transmodern metaphysics introduces is the generality of a dynamical and open background that is overdetermined by the subject's own work motion. In this way the static-fixed background of transcendental ideality, physical laws, and secular power are conceived of as absolute only in relation to the subject's self-positing. This means ultimately that the real in a transmodern sense does not exist before and after the subject's own individuated becoming. As already emphasized the real in a transmodern sense is something that emerges and depends on the symbolic observers enacted transformations in history.

Thus, this transmodern metaphysics can approach technological singularity in an open and incomplete metaphysics to be determined by symbolic observers. In this approach it is posited that totality is nothing but a multiplicity of observers that circumambulate around an absent central unity appearing in an emergent intersubjective space constituted by individuated relations. In other words this is a real truth where each individuated unit of transformative identity is capable of producing an objectively recognized historical real as a pure difference or division via epistemological knowledge structures. These knowledge structures can retroactively change the nature of ontological being itself towards a higher connection or unity (as in the introduction of the symbolic forms of ancient, modern, and postmodern metaphysics throughout historical becoming). Consequently, on the approach to technological singularity, a transmodern metaphysics would predict the breakdown of historical visions of unified totality, and the consequent emergence of ever greater numbers of novel divisions that produce ever greater numbers of novel unities. The totality of divisions may represent the logical consequence of the becoming of the concept in universal history capable of producing a meta-level of unity inaccessible to any one consciousness but which is one consciousness in-itself.

To summarize the transmodern approach to both the historical synthesis of historical forms of totality and the individuated technological singularity let us consider this structure of totality in terms of divisions that introduce new unities. The ancient metaphysical system became a historical real through the pure division of religious and philosophical thinkers in Mesopotamia, Egypt, and Greece which allowed for the formation of a higher unity of transcendental ideality that objectively structured the intersubjective becoming of early agricultural societies for millennia. The modern metaphysical system became a historical real through the pure division of scientific and naturalist thinkers in Europe which allowed for the formation of a higher unity of physical laws that objectively structured the intersubjective becoming of early industrial societies for centuries. The postmodern metaphysical system became a historical real through the pure division of social and philosophical thinkers in Europe and North America which allowed for the formation of a higher unity of deconstructing secular power that objectively structured the intersubjective becoming of early information age societies for decades. Now in the transmodern world in-itself this same force of pure division is predicted to be expressed in increasingly distributed forms (independent of transcendental superspace, physical laws, and secular power) allowing for, perhaps, a higher proliferation of unities than ever before in human history.

In order to connect this notion of metaphysical totality built around emergent subject-object division aiming for unity with cosmic evolutionary philosophy as representing one unified ordering force we may frame it as a potential 'third path' to quantum gravity. Here the nature of a division is a depths where the subject appears as a quantum mechanical entity; and the nature of a unity is a height where the object appears as a gravitationally relativistic entity. In this third path to quantum gravity the real as truth of being is structured by multiple observers each with an attractor dependent horizon that attempts to form an eternal unity in its becoming. Consequently, this path becomes structured by multiple observers inside the universe as opposed to taking a structure of multiple universes as understood by one mythical observer outside the universe, as in quantum multiverse speculations [104]. Thus, instead of attempting to explain the primordial unity of the universe with recourse to a multiverse of infinite physical universes, we need only include the virtual potentiality and actual tendencies of human observers (a multiverse of observers). In this way the primordial astrophysical singularity where all being is one unity could meet technological singularity by way of the demands for unity by an observational multiplicity. Here cosmic evolutionary philosophy as a universal dynamic ideational motion emerges requiring dialectical analysis gives the appearance of a totality structured by an open-incomplete 4-dimensional sphere aiming for closure-completion.

Finally, this transmodern metaphysics as a theory of consciousness may allow us to approach the technological singularity in a novel way by understanding how epistemological constructs of general humanity become a fundamental part of ontological being. In terms of a human subject-oriented approach to technological singularity capable of reconciling our de-centered cosmic position we should emphasize that when we include the future actuality of artificial intelligence, genetic engineering, and quantum computing, we open up a totally new possibility space for observationally constituted dynamical action. In other words, this theory of consciousness requires us to include the future real of knowledge as an activity into the ontology of being as opposed to continuing to focus on an imaginary knowledge that passively reflects the real ontology of being. The consequences of such a perspectival shift forces us to confront the fact that although meaning does exist outside of the symbolic order out in the cosmos, meaning does have a concrete materiality within the symbolic order signaling orientation to higher unity.

In order to work towards being able to think such a reality we should start with the philosophical foundation. The possibility of including the future real of human knowledge as an activity into the fundamental ontology of being was formally opened with modern philosophical idealism and the identification of the a priori conceptual frame as a horizon of being [105]. Towards understanding how this fuller understanding of the relation between human knowledge and natural being itself manifests today we may draw an analogy related to modern physics. In modern physics there is a fundamental shift that has been occurring in high theory from desiring to know 'what the fundamental eternal laws of the physical universe are' to desiring to know 'why does the universe have the particular set of eternal physical laws that it does?' [106]. In order to properly resolve this fundamental shift in ontological questioning we must be capable of a perspective shift within physics itself that appreciates the ontological meaning of quantum computer theory [107], and the consequences of future quantum computation [108]. Here we have a form of fundamental physics knowledge which suggests that observers inside the universe can make an object with their knowledge structures (i.e., supercomputer) that can simulate any physical process (i.e., a physical universe).

The radicality of such a possibility as it relates to technological singularity cannot be understated, but how can we make a division capable of motivating future research in this direction? Here I will make a conjecture that when we are thinking totality from the perspective of subject-object division aimed at unity we need to make sense of fundamental ontological problems related to historical forms of totality. To be specific there appears still unresolved problems in science and mathematics as to both the fundamental natures of mathematical ideality and physical law. From the perspective of ancient metaphysics mathematical ideality exists in a transcendental superspace from eternity; and from the perspective of modern metaphysics physical laws exist in a natural space from eternity. Of course, most contemporary theorists are skeptical of both conjectures, even if both assumptions structure much of science and mathematics. For example, consider that science is often embedded in space and time as universal organizing categories, and mathematics is often perceived to represent a universal knowledge independent of context and history.

In the fundamentally emergentist transmodern totality we may be able to reconcile both problems by positing that mathematical ideality and physical law could be a part of an eternal loop or sphere where physical law emerge (astrophysical singularity; big bang) as a sensual background for observers to construct logical mathematical ideality; and observers constructing logical mathematical ideality emerge as a background capable of constituting sensual physical law (technological singularity). Indeed, is not a fundamental problem in quantum gravity the fact that 'eternal' physical laws of spacetime break down at the singularity of the big bang and the singularity of black holes? In this way, by including the multiplicity of individuating observers approaching technological singularity with their own loops of sense and logic, we may be able to reconcile the breakdown of laws and the constitution of new laws in one transmodern metaphysical system. From this perspective history is fundamentally structured by an observationally constituted expansion of freedom independent of spacetime coordinates that is predicted to result in an objectivity overdetermined by observers (Figure 13).

In this transmodern theory of consciousness, we must be capable of thinking how an individuated observer dividing being with a higher unity [109] could possibly by responsible for the generation of physical law [59] via the immanence of higher order technological possibility spaces constituted by ideational curvature [34,110]. Perhaps this is a way to understand the meaning of quantum computation on the level of fundamental physics and universal history where the physical laws themselves can become radically other via ideal manipulations. Thus, what this division suggests for a higher unity is precisely that researchers interested in understanding totality must take seriously a conscious totality that is divided between subject-object. The reconciliation of such a division requires the emergence of a qualitative phase transition where observers can themselves actively constitute the object-in-question as opposed to merely reflecting given being. In this sense, the dialectical approach to technological singularity is concerned with the way in which historically engaged individuated observers become central to future theories of totality.

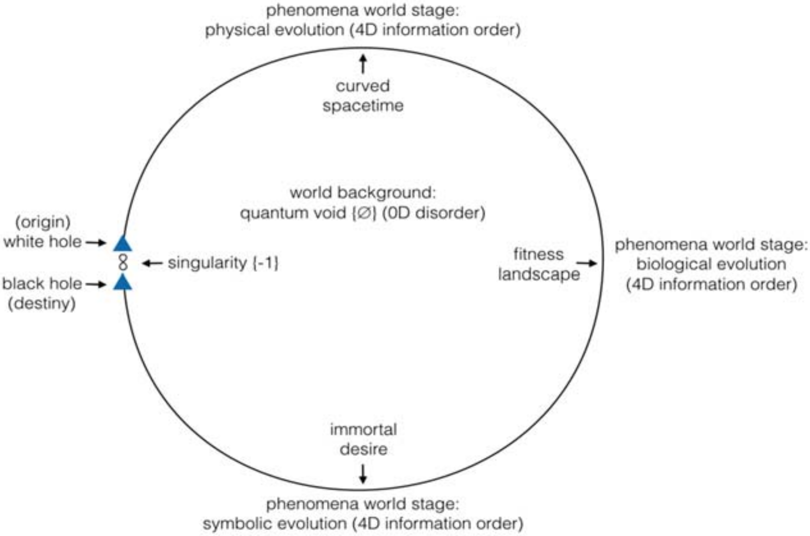


Figure 13. Cosmic evolution and dialectic connecting beginning and end. In this representation, we see the contemporary cosmic evolutionary process that can be divided between physical evolution of curved spacetime, biological evolution of fitness landscapes, and symbolic evolution of immortal desire. Here, although highly speculative, we can start to entertain a potential theoretical link between the mystery of the ordered astrophysical singularity at the beginning of spacetime and the mystery of the destiny of ordered evolutionary processes. From this link, we are asked to think the way in which the horizon of ideation can be possibly inscribed into the immanence of cosmic-physical processes via the inclusion of a type of extreme primordial and emergent curvature where evolutionary change can be reconciled with eternity. The true question here becomes the ultimate nature of the 'eternity'. Here, it is negativized (-1) as absent in the historical process since its presence would nihilate our 4D existence. However, the possibility of the actualizing and perceiving higher super-symmetrical dimensions is mathematically real and experientially realizable given the known future technological possibility space available to future observers.

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

1. Heylighen, F. Complexity and Evolution: Fundamental Concepts of a New Scientific Worldview. Lecture Notes 2014–15. 2014, p. 30. Available online: http://pespmc1.vub.ac.be/books/Complexity-Evolution.pdf (accessed on 31 May 2017).
2. Christian, D. What is Big History? *J. Big Hist.* **2017**, *1*, 4–19. [CrossRef]
3. Polchinski, J. *String Theory: Volume 2, Superstring Theory and Beyond*; Cambridge University Press: Cambridge, UK, 1998.
4. Russell, S.J.; Norvig, P. *Artificial Intelligence: A Modern Approach*; Prentice-Hall: Upper Saddle River, NJ, USA, 1995.
5. Carr, P.A.; Church, G.M. Genome engineering. *Nat. Biotechnol.* **2009**, *27*, 1151–1162. [CrossRef] [PubMed]
6. Last, C. Big Historical Foundations for Deep Future Speculations: Cosmic Evolution, Atechnogenesis, and Technocultural Civilization. *Found. Sci.* **2017**, *22*, 39–124. [CrossRef]
7. Verelst, K. Concerning the ontology of the World-Tree. In *Visions on Nature—Comparative Studies on the Theory of Gaia and Culture*; Elders, F., Ed.; VUB Press: Brussel, Belgium, 2004; pp. 96–122.
8. Sloterdijk, P. *Spheres. Volume 1: Bubbles (Microspherology)*; Semiotext(e): Los Angeles, CA, USA, 2011.

9. Weinert, F. *Copernicus, Darwin and Freud: Revolutions in the History and Philosophy of Science*; Wiley-Blackwell: Hoboken, NJ, USA, 2009.
10. Vidal, C. *The Beginning and the End: The Meaning of Life in a Cosmological Context*; Springer: Berlin, Germany, 2014.
11. Kojève, A. *Introduction to the Reading of Hegel: Lectures on the Phenomenology of Spirit*; Cornell University Press: Ithaca, NY, USA, 1980.
12. Koyré, A. *From the Closed World to the Infinite Universe*; Library of Alexandria: Alexandria Governorate, Egypt, 1957.
13. Turchin, V. *The Phenomenon of Science: A Cybernetic Approach to Human Evolution*; Columbia University Press: New York, NY, USA, 1977.
14. Vinge, V. The coming technological singularity. *Whole Earth Rev.* **1993**, *81*, 88–95.
15. More, M.; Vita-More, N. Part VIII—Future Trajectories Singularity. In *The Transhumanist Reader*; More, M., Vita-More, N., Eds.; Wiley-Blackwell: Hoboken, NJ, USA, 2013; pp. 361–363.
16. Vinge, V. Technological Singularity. In *The Transhumanist Reader*; More, M., Vita-More, N., Eds.; Wiley-Blackwell: Hoboken, NJ, USA, 2013; pp. 365–375.
17. Soares, N.; Fallenstein, B. Agent Foundations for Aligning Machine Intelligence with Human Interests: A Technical Research Agenda. In *The Technological Singularity: Managing the Journey*; Callaghan, V., Miller, J., Yampolskiy, R., Armstrong, S., Eds.; Springer: Berlin, Germany, 2017; pp. 103–126.
18. Moore, G. Cramming more components onto integrated circuits. *Electronics* **1965**, *38*, 1–4. [CrossRef]
19. Moore, G. Progress in digital integrated electronics. Available online: https://www.eng.auburn.edu/~agrawvd/COURSE/E7770_Spr07/READ/Gordon_Moore_1975_Speech.pdf (accessed on 5 April 2018).
20. Kurzweil, R. The law of accelerating returns. Available online: http://www.kurzweilai.net/the-law-of-accelerating-returns (accessed on 5 April 2018).
21. Kurzweil, R. *The Singularity Is Near: When Humans Transcend Biology*; Penguin: London, UK, 2005.
22. Sirius, R.U.; Cornell, J. *Transcendence: The Disinformation Encyclopedia of Transhumanism and the Singularity*; Disinformation Books: New York, NY, USA, 2015.
23. Hanson, R. *The Age of Em: Work, Love and Life When Robots Rule the Earth*; Oxford University Press: Oxford, UK, 2016.
24. Goertzel, B.; Goertzel, T. (Eds.) The End of the Beginning: Life, Society, and Economy on the Brink of Singularity. Humanity + Press: San Jose, CA, USA, 2015.
25. Ulam, S. Tribe to John von Neumann. *Bull. Am. Math. Soc.* **1958**, *64*, 1–49. [CrossRef]
26. Good, I.J. Speculations concerning the first ultraintelligent machine. *Adv. Comput.* **1965**, *6*, 31–83.
27. Lenartowicz, M. Creatures of the semiosphere: A problematic third party in the 'humans plus technology' cognitive architecture of the future global superintelligence. *Technol. Forecast. Soc. Chang.* **2017**, *114*, 35–42. [CrossRef]
28. Zalasiewicz, J.; Williams, M.; Smith, A.; Barry, T.L.; Coe, A.L.; Bown, P.R.; Brenchley, P.; Cantrill, D.; Gale, A.; Gibbard, P.; et al. Are we now living in the Anthropocene? *GSA Today* **2008**, *18*, 4–8. [CrossRef]
29. Wark, M. *General Intellects: Twenty-One Thinkers for the Twenty-First Century*; Verso: London, UK, 2017.
30. Barrat, J. *Our Final Invention: Artificial Intelligence and the End of the Human Era*; St. Martin's Press: New York, NY, USA, 2013.
31. Bostrom, N. *Superintelligence: Paths, Dangers, Strategies*; Oxford University Press: Oxford, UK, 2014.
32. Nicolelis, M.; Cicurel, R. *The Relativistic Brain: How It Works and Why It Cannot be Simulated by a Turing Machine*; Kios Press: Sao Paulo, Brazil, 2015.
33. Kaku, M. *The Future of the Mind: The Scientific Quest to Understand and Empower the Mind*; Anchor Books: New York, NY, USA, 2014.
34. Smith, H.; Marranca, R. *The World's Religions*; Harper One: New York, NY, USA, 2009.
35. Wolfson, H.A.; Fackenheim, E.L. Philo: Foundations of Religious Philosophy in Judaism, Christianity, and Islam. *Rev. Metaphys.* **1947**, *1*, 89.
36. Mercer, C.; Trothen, T.J. (Eds.) *Religion and Transhumanism: The Unknown Future of Human Enhancement*; Praeger: London, UK, 2015.
37. Hughes, J. The politics of transhumanism and the techno-millennial imagination, 1626–2030. *Zygon* **2012**, *47*, 757–776. [CrossRef]

38. Wolyniak, J. "The Relief of Man's Estate": Transhumanism, the Baconian Project, and the Theological Impetus for Material Salvation. In *Religion and Transhumanism: The Unknown Future of Human Enhancement*; Mercer, C., Trothen, T.J., Eds.; Praeger: London, UK, 2015; pp. 53–70.
39. Pross, A.; Pascal, R. The origin of life: What we know, what we can know, and what we will never know. *Open Biol.* **2013**, *3*, 120190. [CrossRef] [PubMed]
40. Thompson, E. *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*; Harvard University Press: Cambridge, MA, USA, 2002.
41. Dunbar, R. Why only humans have language. In *The Prehistory of Language*; Botcha, R., Knight, C., Eds.; Oxford University Press: Oxford, UK, 2009; pp. 1–26.
42. Chesterton, G.K. The Everlasting Man. 1925. Available online: http://www.gkc.org.uk/gkc/books/everlasting_man.pdf (accessed on 5 January 2018).
43. Hegel, G.W.F. *Phenomenology of Spirit*; Miller, A.V.; Findlay, J.N., Translators; Motilal Banarsidass Publishers: Delhi, India, 1998.
44. Christian, D. *Maps of Time: An Introduction to Big History*; Berkeley University Press: Berkeley, CA, USA, 2004.
45. Kurzweil, R. *The Singularity Is Near: When Humans Transcend Biology*; Penguin: London, UK, 2005; p. 28.
46. Stewart, J. The Meaning of Life in a Developing Universe. *Found. Sci.* **2010**, *15*, 395–409. [CrossRef]
47. Heylighen, F. Self-organization of complex, intelligent systems. Available online: http://134.184.131.111/papers/ECCO-paradigm.pdf (accessed on 5 January 2018).
48. Spier, F. *Big History and the Future of Humanity*; Wiley Blackwell: Hoboken, NJ, USA, 2015.
49. Spier, F. How big history works: Energy flows and the rise and demise of complexity. *Soc. Evol. Hist.* **2005**, *4*, 87–135.
50. Chaisson, E. *Cosmic Evolution: The Rise of Complexity in Nature*; Harvard University Press: Cambridge, MA, USA, 2001.
51. Smart, J. Evo Devo Universe? A Framework for Speculations on Cosmic Culture. In *Cosmos and Culture: Cultural Evolution in a Cosmic Context*; Dick, S.J., Lupisella, M.L., Eds.; Govt Printing Office, NASA SP-2009-4802; Govt Printing Office: Washington, DC, USA, 2008.
52. Stewart, J. The direction of evolution: The rise of cooperative organization. *Biosystems* **2014**, *123*, 27–36. [CrossRef] [PubMed]
53. Kauffman, S.A. The origins of order: Self-organization and selection in evolution. In *Spin Glasses and Biology*; World Scientific: Singapore, 1992; pp. 61–100.
54. Heylighen, F.; Beigi, S.; Veloz, T. Chemical Organization Theory as a modelling framework for self-organization, autopoiesis and resilience. Available online: https://pdfs.semanticscholar.org/e9ff/09b77597f8c0499a3ef4de126e29687cc713.pdf (accessed on 2 January 2018).
55. Veitas, V.; Weinbaum, D. A world of views: A world of interacting post-human intelligences. In *The Beginning and the End: Life, Society, and Economy on the Brink of Singularity*; Goertzel, B., Goertzel, T., Eds.; Humanity + Press: San Jose, CA, USA, 2015; pp. 495–567.
56. Vidal, C. *The Beginning and the End: The Meaning of Life in a Cosmological Context*; Springer: Berlin, Germany, 2014; p. ix.
57. Krauss, L. *A Universe from Nothing: Why There Is Something Rather than Nothing*; Free Press: New York, NY, USA, 2012.
58. Aunger, R. Major transitions in 'big history'. *Technol. Forecast. Soc. Chang.* **2007**, *74*, 1137–1163. [CrossRef]
59. Von Bertalanffy, L. *General System Theory: Foundation, Development, Applications*; G. Braziller: New York, NY, USA, 1968.
60. Weaver, W. Science and Complexity. *Am. Sci.* **1948**, *36*, 536–544. [PubMed]
61. Heylighen, F. Complexity and Evolution: Fundamental Concepts of a New Scientific Worldview. Lecture Notes 2014–15. 2014, pp. 55–57. Available online: http://pespmc1.vub.ac.be/books/Complexity-Evolution.pdf (accessed on 31 May 2017).
62. Rovelli, C.; Vidotto, F. *Covariant Loop Quantum Gravity: An Elementary Introduction to Quantum Gravity and Spinfoam Theory*; Cambridge University Press: Cambridge, UK, 2015.
63. Latour, B. *An Introduction to Actor-Network Theory*; Oxford University Press: Oxford, UK, 2005.
64. DeLanda, M. *Intensive Science and Virtual Philosophy*; Bloomsbury: Hong Kong, China, 2013; p. 59.
65. Carroll, B.W.; Ostlie, D.A. *An Introduction to Modern Astrophysics*; Cambridge University Press: Cambridge, UK, 2017.

66. Corning, P. The Re-Emergence of "Emergence". *Complexity* **2002**, *7*, 18–30. [CrossRef]
67. Lineweaver, C.H.; Davies, P.C.W.; Ruse, M. *Complexity and the Arrow of Time*; Cambridge University Press: Cambridge, UK, 2013.
68. Smolin, L. *Time Reborn*; Houghton Mifflin Harcourt: Boston, MA, USA, 2013.
69. Zeh, H.D. *The Physical Basis of the Direction of Time*; Springer: New York, NY, USA, 1990.
70. Aunger, R. A rigorous periodization of 'big' history. *Technol. Forecast. Soc. Chang.* **2007**, *74*, 1164–1178. [CrossRef]
71. Prigogine, I.; Stengers, I. *Order out of Chaos: Man's New Dialogue with Nature*; Bantam Books: New York, NY, USA, 1984.
72. DeLanda, M. *Intensive Science and Virtual Philosophy*; Bloomsbury: Hong Kong, China, 2013; p. 65.
73. Weinbaum, D.R. Complexity and the Philosophy of Becoming. *Found. Sci.* **2015**, *20*, 283–322. [CrossRef]
74. Žižek, S. *Less Than Nothing: Hegel and the Shadow of Dialectical Materialism*; Verso: London, UK, 2012.
75. Heylighen, F. Return to Eden? Promises and perils on the road to global superintelligence. In *The Beginning and the End: Life, Society, and Economy on the Brink of Singularity*; Goertzel, B., Goertzel, T., Eds.; Humanity + Press: San Jose, CA, USA, 2015.
76. Kyriazis, M. Systems neuroscience in focus: From the human brain to the global brain? *Front. Syst. Neurosci.* **2015**, *9*. [CrossRef] [PubMed]
77. Husserl, E. *The Crisis of European Sciences and Transcendental Phenomenology: An Introduction to Phenomenological Philosophy*; Northwestern University Press: Evanston, IL, USA, 1970.
78. Hawkins, J.; Blakeslee, S. *On Intelligence: How a New Understanding of the Brain Will Lead to the Creation of Truly Intelligent Machines*; Macmillan: Basingstoke, UK, 2007.
79. Heylighen, F. Life Is an Adventure: An Agent-Based Reconciliation of Narrative and Scientific Worldviews. Available online: http://pespmc1.vub.ac.be/Papers/Life-Adventure.pdf (accessed on 4 October 2017).
80. Adams, F.; Laughlin, G. *The Five Ages of the Universe: Inside the Physics of Eternity*; The Free Press: New York, NY, USA, 1999.
81. Deacon, T. *Incomplete Nature: How Mind Emerged from Matter*; W.W. Norton and Company: New York, NY, USA, 2011.
82. Žižek, S. *Less Than Nothing: Hegel and the Shadow of Dialectical Materialism*; Verso: London, UK, 2012; pp. 3–5.
83. Lenartowicz, M.; Weinbaum, D.R.; Braathen, P. Social systems: Complex adaptive loci of cognition. *Emerg. Complex. Organ.* **2016**, *18*. [CrossRef]
84. Plato, R.E. *The Dialogues of Plato: Volume II: The Symposium*; Translated with Comment by Allen, R.E.; Yale University Press: London, UK, 1998.
85. Lacan, J. Knowledge and truth. In *The Seminar of Jacques Lacan. Book XX: On Feminine Sexuality, The Limits of Love and Knowledge, 1972–73*; Jacques-Alain, M., Ed.; Translated with Notes by Bruce Fink; W.W. Norton and Company: New York, NY, USA, 1999; p. 90.
86. Žižek, S. *Absolute Recoil: Towards a New Foundation of Dialectical Materialism*; Verso: London, UK, 2014; p. 255.
87. Belfiore, F. *The Triadic Structure of the Mind: Outlines of a Philosophical System*; University Press of America: Lanham, MD, USA, 2014.
88. Stasheff, J. From operads to 'physically' inspired theories. In *Operads: Proceedings of Renaissance Conferences*; American Mathematical Soc.: Providence, RI, USA, 1997; Volume 202, pp. 53–82.
89. Forcey, S.; Springfield, D. Geometric combinatorial algebras: Cyclohedron and simplex. *J. Algebraic Comb.* **2010**, *32*, 597–627. [CrossRef]
90. Evans, D. *An Introductory Dictionary of Lacanian Psychoanalysis*; Routledge: London, UK, 2006; p. 1.
91. Peterson, J.B. *Maps of Meaning: The Architecture of Belief*; Routledge: New York, NY, USA, 1999; pp. 332,334.
92. Žižek, S. *Less Than Nothing: Hegel and the Shadow of Dialectical Materialism*; Verso: London, UK, 2012; p. 17.
93. Heylighen, F. Complexity and Evolution: Fundamental Concepts of a New Scientific Worldview. Lecture Notes 2014–15. 2014, p. 139. Available online: http://pespmc1.vub.ac.be/books/Complexity-Evolution.pdf (accessed on 31 May 2017).
94. Hegel, G.W.F. *The Science of Logic*; Cambridge University Press: Cambridge, UK, 2010.
95. Kurzweil, R. How my predictions are faring. *Kurzweil AI* **2010**, 1–146.
96. Kurzweil, R. *How to Create a Mind: The Secret of Human thought Revealed*; Penguin: New York, NY, USA, 2012.
97. Jameson, F. *The Valences of the Dialectic*; London: Verso, 2009.
98. Evans, D. *An Introductory Dictionary of Lacanian Psychoanalysis*; Routledge: London, UK, 2006; p. 187.

99. Smolin, L. *Three Roads to Quantum Gravity*; Basic Books: New York, NY, USA, 2001; pp. 9–10.
100. Penrose, R. *The Road to Reality: A Complete Guide to the Laws of the Universe*; A.A. Knopf: New York, NY, USA, 2004; p. 1018.
101. Frolov, V.P.; Zelnikov, A. *Introduction to Black Hole Physics*; Oxford University Press: Oxford, UK, 2011; p. 340.
102. Zagzebski, L. The Two Greatest Ideas. KU Leuven University, Centre for Logic and Analytic Philosophy, 2017. Available online: https://hiw.kuleuven.be/claw/events/agenda/kardinaal-mercier-lecture-linda-zagzebski-the-two-greatest-ideas (accessed on 22 October 2017).
103. Smolin, L. *Three Roads to Quantum Gravity*; Basic Books: New York, NY, USA, 2001; p. 10.
104. Smolin, L. *Three Roads to Quantum Gravity*; Basic Books: New York, NY, USA, 2001; p. 48.
105. Žižek, S. *Less Than Nothing: Hegel and the Shadow of Dialectical Materialism*; Verso: London, UK, 2012; p. 9.
106. Smolin, L. *Three Roads to Quantum Gravity*; Basic Books: New York, NY, USA, 2001.
107. Deutsch, D. Quantum theory, the Church-Turing principle and the universal quantum computer. *Proc. R. Soc. Lond. A Math. Phys. Eng. Sci.* **1985**, *400*, 97–117. [CrossRef]
108. Lloyd, S. *Programming the Universe: A Quantum Computer Scientist Takes on the Cosmos*; Knopf: New York, NY, USA, 2006.
109. Penrose, R. *The Road to Reality: A Complete Guide to the Laws of the Universe*; A.A. Knopf: New York, NY, USA, 2004.
110. Drexler, E. *Radical Abundance: How a Revolution in Nanotechnology Will Change Civilization*; Public Affairs: New York, NY, USA, 2013.

information

MDPI

Article

Can Computers Become Conscious, an Essential Condition for the Singularity?

Robert K. Logan

Department of Physics, University of Toronto, 60 St. George, Toronto, ON M5S 1A7, Canada;
logan@physics.utoronto.ca; Tel.: +1-(416)-978-8632

Received: 12 November 2017; Accepted: 6 December 2017; Published: 9 December 2017

Abstract: Given that consciousness is an essential ingredient for achieving Singularity, the notion that an Artificial General Intelligence device can exceed the intelligence of a human, namely, the question of whether a computer can achieve consciousness, is explored. Given that consciousness is being aware of one's perceptions and/or of one's thoughts, it is claimed that computers cannot experience consciousness. Given that it has no sensorium, it cannot have perceptions. In terms of being aware of its thoughts it is argued that being aware of one's thoughts is basically listening to one's own internal speech. A computer has no emotions, and hence, no desire to communicate, and without the ability, and/or desire to communicate, it has no internal voice to listen to and hence cannot be aware of its thoughts. In fact, it has no thoughts, because it has no sense of self and thinking is about preserving one's self. Emotions have a positive effect on the reasoning powers of humans, and therefore, the computer's lack of emotions is another reason for why computers could never achieve the level of intelligence that a human can, at least, at the current level of the development of computer technology.

Keywords: computers; consciousness; singularity; artificial general intelligence; intelligence; emotion; self

1. Introduction: Thinking, Language and Communication

Advocates of technological Singularity believe that through a process of iteration of artificial general intelligence (AGI), one day, an AGI device will design a computer or a robot with greater intelligence that it, and that the computer will do the same until a computer will be created with an intelligence greater than that of any human being. Because a robot is basically a computer with moving mechanical parts, the term computer is used henceforth to refer to both computers and robots.

In order for this Singularity to be achieved, the AGI device will have to achieve consciousness that includes being aware of what it knows. The purpose of this essay is to show that this is not possible, and hence, that the idea of Singularity is a pipe dream.

Before embarking on this quest, we want the reader to understand that a distinction will be made between the brain and the mind. All vertebrates have a brain, but humans are the only species that have a mind that evolved from their brain, which has the added features of verbal language, mathematical thinking, the ability to plan and the ability to conceive of things that are not immediately available in the here and now. Before humans achieved verbal language, their brain was basically a percept processor. With verbal language, the brain evolved into a mind capable of conceptual thinking. I believe that thought, language and communication are interconnected or hyperlinked. Many linguists regard language as the medium by which we communicate our thoughts. I believe that language is also the medium by which we formulate our conceptual thinking. I regard thinking as silent language and that language also has the additional feature of facilitating the communication of our thoughts.

> The origins of speech and the human mind . . . emerged simultaneously as the bifurcation from percepts to concepts and a response to the chaos associated with the information overload that resulted from the increased complexity of hominid life. As our ancestors developed tool making, controlled fire, lived in larger social groups and engaged in large-scale coordinated hunting their brains could no longer cope with the richness of life solely on the basis of its perceptual sensorium and as a result a new level of order emerged in the form of conceptualization and speech. Speech arose simultaneously as a way to control information and as a medium for communication. Rather than regarding speech as vocalized thought one may just as well regard thought as silent speech. ([1], p. 5)

Concept-based language and thinking emerged because:

> Percepts no longer had the richness or the variety with which to represent and model hominid experience once the new skills of hominids such as tool making, the control of fire and social organization were acquired. It was in this climate that speech emerged and the transition or bifurcation from perceptual thinking to conceptual thinking occurred. The initial concepts were, in fact, the very first words of spoken language. Each word served as a metaphor and strange attractor uniting all of the pre-existing percepts associated with that word in terms of a single word and, hence, a single concept. All of one's experiences and perceptions of water, the water we drink, bathe with, cook with, swim in, that falls as rain, that melts from snow, were all captured with a single word, water, which also represents the simple concept of water. ([1], p. 49)

Thinking, communicating and language form an emergent supervenient system for humans and would be a necessary attribute for any form of intelligence equal to or greater than that of human intelligence.

2. Perception-Based Versus Concept-Based Consciousness

Next, we turn to Ned Block's ([2], p. 227 and 230) observation that there are basically two levels of consciousness that he defined as *phenomenal* consciousness or p-consciousness and *access* consciousness or a-consciousness:

> Phenomenal consciousness is experience; the phenomenally conscious aspect of a state is what it is like to be in that state. The mark of access consciousness [a-consciousness], by contrast, is availability for use in reasoning and rationally guiding speech and action.
>
> P-conscious states are experiential, that is, a state is p-conscious if it has experiential properties. The totality of the experiential properties of a state are 'what it is like' to have it. Moving from synonyms to examples, we have p-conscious states when we see, hear, smell, taste, and have pains.

P-consciousness is perception-based and includes visual, auditory, olfactory, gustatory, tactile, pain, thermoperception, kinesthetic, chemical, magnetic, and equilibriception forms of consciousness. Each form of p-consciousness corresponds to its respective sensory channel or capability and hence is a perception-based consciousness. It is about perceiving or sensing signals either within the subject or those emanating from the environment or umwelt in which the subject operates.

A-consciousness, which is available for reasoning and rationality, is therefore concept-based as reasoning and rationality are concept-based mental activities. A-consciousness or access consciousness is therefore being aware of our thoughts and knowing what we know and hence entails listening to our silent speech. Unless we translate our percepts or the products of our p-consciousness into concepts or language they are not accessible "for use in reasoning and rationally guiding speech and action" and as such will not be part of our a-consciousness.

A-consciousness is closely tied to verbal language as described above. A-consciousness or concept-based consciousness is strictly restricted to human beings because we are the only organism capable of verbal language and hence conceptualization. A-consciousness is basically being aware of our thoughts and knowing what we know and hence is basically listening to ourselves silently talking to ourselves. With this in mind and for the purposes of this discussion I prefer to regard Block's a-consciousness as concept-based and p-consciousness as perception-based consciousness respectively.

3. Why It Is Impossible for a Computer to Possess Consciousness

We can immediately dismiss p- or perception-based consciousness as a possibility for computers as they have no nervous system and an integrated set of sense organs and therefore they cannot perceive at the level of a human. Let us therefore immediately consider whether it is possible for computers to possess a- or concept-based consciousness.

A- or concept-based consciousness, as I have claimed, is being aware of one's thoughts and therefore is basically listening to one's own internal speech. But to have internal speech one must possess external speech, and that requires having a desire or a purpose to communicate. Wanting to communicate, in turn, requires being aware of other intelligences and having a desire to communicate with them with language at least at the level of human language. But the desire to communicate grows out of social needs that requires having emotions of which the computers have none. Emotions arise from the physical interactions of a living organism initiated by sensory input.

Robert Worden [3] attributes primate social skills to the development of human language by proposing that "language is an outgrowth of primate social intelligence". One of his key hypotheses is that: "The internal representation of language meaning in the brain derives from the primate representation of social situations .... While some use of language is internal, for thought processes, this suggests strongly that it is an outgrowth of social intelligence ([3], p. 153)". Computers, on the other hand, have no social skills as social skills are based on emotions. The emotions of love, caring friendship and altruism are adaptive and increase the survival rate of organisms that possess them. Computers are not organisms, they have no will to live, they have no reason to communicate. They have no need to be adaptive.

The desire to communicate verbally has been attributed to three closely related attributes of human cognition, namely, a theory of mind, the sharing of joint attention, and the advent of altruistic behavior. In order to want to engage in the joint attention that Tomasello ([4], pp. 208–209) suggests was essential for the emergence of language it is necessary to have a theory of mind ([5], p. 102), namely the realization that other humans have a mind, desires and needs similar to one's own mind, desires and needs. At the same time, there must have developed a spirit of altruism ([6], p. 41) once a theory of mind emerged so that human conspecifics would want to enter into the cooperative behavior that is entailed in the sharing of information. Theory of mind and joint attention catalyzes the social function of communication and cooperative behavior and vice-versa. As computers could not have a theory of mind, have no reason to be altruistic and joint attention cannot take place in real time they have no desire to initiate communication and do not communicate unless the communication is initiated by their users.

Emotions, communications, and language are all interlinked as Darwin [7] pointed out long ago in his book, *The Expression of the Emotions in Man and Animal*, as noted by Hess and Thibault ([8], p. 120):

> Darwin's basic message was that emotion expressions are evolved and (at least at some point in the past) adaptive. For Darwin, emotion expressions not only originated as part of an emotion process that protected the organism or prepared it for action but also had an important communicative function. Darwin ([7], p. 368) saw in this communicative function a further adaptive value when he wrote: "We have also seen that expression in itself, or the language of the emotions, as it has sometimes been called, is certainly of importance for the welfare of mankind".

If language emerged from the social skills of primates, which are emotion based, then it is hard to conceive how computers could evolve language as they have no emotions. I would add that since having social skills and the desire to communicate requires having emotions that in turn requires being alive. I cannot imagine how a computer could develop language. And if computers could not develop language how would they be able to have concepts and hence concept based consciousness. I therefore conclude that those that want to create the Singularity will have to figure out how to create a living creature from scratch with the complexity and emotions of a human, something that biologists cannot even imagine.

The AGI computer, that believers in the Singularity think will be created some day, will have to be capable of saying spontaneously without being programmed something along the lines of "I think; therefore, I am," just as Descartes did when he said in Latin, "cogito ergo sum" and in colloquial French "je pense, donc je suis".

Language not only allows humans to communicate abstract concepts to each other but it is also put to use for the internal dialogue of conceptual thinking. In my book The Extended Mind: The Emergence of Language, the Human Mind and Culture [1] I proposed that the mind is more than just the brain and that with language the mind was able to conceptualize in addition to processing percepts. As a consequence the human mind emerged with verbal language so that the mind = the brain plus language.

This idea parallels Darwin's [7] expression of the co-evolution of language and the intellectual power of humans. It can be found in Chapter 21, p. 92 of *The Descent of Man*: "A complex train of thought can no more be carried on without the aid of words, whether spoken or silent, than a long calculation without the use of figures or algebra".

A computer through AGI can become a brain of sorts but not a mind because it does not possess language and therefore cannot listen to its internal speech and therefore cannot become conscious. A form of intelligence that is not conscious of its mental processes is severely limited and therefore could never compete with the human mind.

4. AI Does Not Take into Account Work or Biology: An Acknowledgement

One of the core ideas in the argument I have presented in this essay is that the Singularity is not possible because computers are not and cannot be living organisms. While working on this project and searching my computer I encountered some notes I took of a conversation I had with Stuart Kauffman in 2006 while we were working on a paper entitled Propagating Organization: An Inquiry [9] (Kauffman et al. 2007). I share these notes because they reinforce my position that silicon based AI can never duplicate human intelligence giving credit to Stuart for what is of value and accept responsibility for whatever this argument lacks:

> It takes work to make information. Information is embodied in some specific pattern of matter and energy. It takes work to pattern or shape that matter and energy. Life is a shaper of matter and energy that is capable of doing a work cycle if free energy is available. **AI does not take into account work or biology** (I bolded the relevant part of these notes for this essay).

5. Robots and the Singularity

We have used the term computer to represent either a computer or a computerized robot. Murray Shannahan ([10], p. 5) has argued that "the only way to achieve human-level AI ... is through robotics" and a robot "with a biomimetic set of sensors ([10], p. 37)". Shanahan suggests that through "whole brain emulation ... produced by scanning a brain and thereby producing a high-fidelity, neuron-for-neuron and synapse-for-synapse simulation ([10], p. 119)" consciousness could be achieved. As intriguing as Shanahan's ideas are they fall short, in the opinion of this author, for the reason they do not into account two things: (1) the Kauffman criteria that a living organism must be capable of doing a work cycle and (2) Intelligence has an emotional component as described in the conclusion section below.

6. Conclusion: Emotions and Reasoning

The lack of emotions severally limits the scope of AI and makes AGI an impossible dream. Damasio [11] study of emotions revealed that:

> Emotion is always in the loop of reason. Emotion is an adaptive response, part of the vital process of normal reasoning and decision-making. It is one of the highest levels of bioregulation for the human organism and has an enormous influence on the maintenance of our homeostatic balance and thus of our well-being. Last but not least, emotion is critical to learning and memory.

Since computers are non-biological they have no emotion and since according to Damasio emotions play an important role in reasoning, decision-making and learning, I believe, that the idea of the Singularity is an impossible dream. I am well aware that it was once proven that heavier than air flight was impossible which is why I have softened my conclusion as a belief. The obstacles to make a computer or a robot conscious are formidable as I have outlined in my article. AI and robotics as tools, however, in partnership with its human creators will advance human knowledge and productivity. Perhaps in an attempt to make computers/robots conscious other useful technology will be discovered so if the reader does not agree with my conclusions they at least have an idea of the scope of the challenge they face.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Logan, R.K. *The Extended Mind: The Emergence of Language, the Human Mind and Culture*; University of Toronto Press: Toronto, ON, Canada, 2007.
2. Block, N. On a Confusion about a Function of Consciousness. *Behav. Brain Sci.* **1995**, *18*, 227–247. [CrossRef]
3. Worden, R.P. The Evolution of Language from Social Intelligence. In *Approaches to the Evolution of Language*; Hurford, J., Studdert-Kennedy, M., Knight, C., Eds.; Cambridge University Press: Cambridge, UK, 1998; pp. 148–168.
4. Tomasello, M. Introduction: A Cognitive-functional Perspective on Language Structure. In *The New Psychology of Language: Cognitive-Functional Perspective on Language Structure*; Tomasello, M., Ed.; Erlbaum: Mahwah, NJ, USA, 1998; Volume 1, pp. vii–xxi.
5. Dunbar, R. Theory of Mind and the Evolution of Language. In *Approaches to the Evolution of Language*; Hurford, J.R., Studdert-Kennedy, M., Knight, C., Eds.; Cambridge University Press: Cambridge, UK, 1998; pp. 92–110.
6. Ulbaek, I. The origin of language and cognition. In *Approaches to the Evolution of Language*; Hurford, J.R., Studdert-Kennedy, M., Knight, C., Eds.; Cambridge University Press: Cambridge, UK, 1998; pp. 30–43.
7. Darwin, C. *The Descent of Man, and Selection in Relation to Sex*; J. Murray: London, UK, 1871; (Reissued in facsimile. Princeton University Press: Princeton, NJ, USA, 1981).
8. Hess, U.; Thibault, P. Darwin and emotion expression. *Am. Psychol.* **2009**, *64*, 120–128. [CrossRef] [PubMed]
9. Kauffman, S.; Logan, R.K.; Este, R.; Goebel, R.; Hobill, D.; Smulevich, I. Propagating Organization: An Enquiry. *Biol. Philos.* **2007**, *23*, 27–45. [CrossRef]
10. Shanahan, M. *The Technological Singularity*; MIT Press: Cambridge, MA, USA, 2015.
11. Damasio, A. The Science of Education. 1998. Available online: http://www.loc.gov/loc/brain/emotion/Damasio.html (accessed on 9 November 2017).

information

MDPI

Essay

The Universality of Experiential Consciousness

Robert K. Logan

Department of Physics, University of Toronto, Toronto, ON M5S 1A, Canada; logan@physics.utoronto.ca

Received: 16 November 2018; Accepted: 11 January 2019; Published: 17 January 2019

Abstract: It is argued that of Block's (On a confusion about a function of consciousness, 1995; The Nature of Consciousness: Philosophical Debates, 1997) two types of consciousness, namely *phenomenal* consciousness (p-consciousness) and *access* consciousness (a-consciousness), that p-consciousness applies to all living things but that a-consciousness is uniquely human. This differs from Block's assertion that a-consciousness also applies to some non-human organisms. It is suggested that p-consciousness, awareness, experience and perception are basically equivalent and that human consciousness has in addition to percept-based p-consciousness, concept-based a-consciousness, a verbal and conceptual form of consciousness that can be utilized to coordinate, organize and plan activities for rational decision-making. This argument is based on Logan's (The Extended Mind: The Emergence of Language, The Human Mind and Culture, 1997) assertion that humans are uniquely capable of reasoning and rationality because they are uniquely capable of verbal language and hence the ability to conceptualize.

Keywords: consciousness; experience; phenomenal consciousness; access consciousness; percept; concept; language

---

Consciousness is a mongrel concept—Ned Block [1]
Sentience is consciousness—Eugene T. Gendlin [2]

Anything that we are aware of at a given moment forms part of our consciousness, making conscious experience at once the most familiar and most mysterious aspect of our lives—Susan Schneider and Max Velmans [3].

1. Introduction

As Schneider and Velmans [3] have stated, consciousness is something we are constantly aware of but find difficult to understand or explain. The perceived mystery of consciousness has been so great that many in the past and some today have denied its existence and considered the study of it a fool's errand. In more recent times this skepticism has abated and it is considered to be a legitimate area of study by many researchers in philosophy, psychology, cognitive science and neuroscience.

The mystery associated with consciousness is finding a way to explain what makes it possible as well as providing answers to the following questions: How is it connected to language and thought? Are non-human animals conscious and to what extent? What is the relationship of consciousness to awareness, perception, experience, conceptualization and thought?

2. Block's Phenomenal Consciousness and Access Consciousness

Ned Block [1] has suggested that there are two types of consciousness, namely, *phenomenal* consciousness (p-consciousness) and *access* consciousness (a-consciousness).

- Phenomenal consciousness is experience; the phenomenally conscious aspect of a state is what it is like to be in that state. The mark of access-consciousness, by contrast, is availability for use in reasoning and rationally guiding speech and action [1] (p. 227).

- P-conscious states are experiential, that is, a state is p-conscious if it has experiential properties. The totality of the experiential properties of a state are 'what it is like' to have it. Moving from synonyms to examples, we have p-conscious states when we see, hear, smell, taste, and have pains [4].
- I agree with Block that all living organisms possess some level of p-consciousness in that they are aware of their environment and experience their surrounding through their channels of perception. Where I disagree with Block is that a-consciousness can occur with non-human organisms. In particular I take issue with the following assertion that Block makes:

A state is access- conscious (A-conscious) if, in virtue of one's having the state, a representation of its content is (1) inferentially promiscuous (Stich 1978), that is, poised for use as a premise in reasoning, (2) poised for rational control of action, and (3) poised for rational control of speech. (I will speak of both states and their contents as A-conscious.) These three conditions are together sufficient, but not all necessary. I regard (3) as not necessary (and not independent of the others), because I want to allow that non- linguistic animals, for example chimps, have A-conscious states. I see A-consciousness as a cluster concept, in which (3) - roughly, reportability - is the element of the cluster with the smallest weight, though (3) is often the best practical guide to A-consciousness [1], [bolding is mine].

I believe there is value in Block's formulation of a- and p- consciousness but I also believe that conceptualization is intimately connected to and dependent on verbal symbolic language, a position I share with others including Brandom [5,6], Davidson [7] and Dummett [8]. The issue of whether or not conceptualization requires language and whether non-human animals are capable of conceptualization is not resolved and perhaps cannot be resolved scientifically because the falsification of the proposition is not possible since we cannot get into the minds of non-human animals as we can with fellow humans through our theory of mind. A theory of mind depends on the assumption that other humans think as I do and that they, like me, have thoughts, reasonings, interests, desires, and intentions similar to mine. My theory of mind is based on my social interactions with other humans and is more precise with those with whom I share a common language and culture than with those from a different language and cultural group. There are limits to my theory of mind as I am not always able to predict the reactions to what I say and do with others even with members of my own family. It is also certainly the case that I and other fellow humans have no theory of mind of non-human animals nor can we. It is, therefore, impossible to know how they think or even if their behavior is guided by thought.

Certainly a-consciousness cannot be available "for use in reasoning and rationally guiding speech" for non-humans as only humans are capable of speech. As for its availability "for use in reasoning and rationally guiding . . . action" for non-human beings, I believe that this is not possible because reasoning and rationality depends on the ability to conceptualize. I have argued elsewhere that conceptualization, the basis for reasoning and rationality depends on the possession of verbal language [4], a unique feature of humans. I will also argue that Block's p-consciousness is percept-based as it is experiential and that a-consciousness is concept-based as it is associated with reasoning and rationality.

3. The Many Levels of Experiential or Phenomenal Consciousness (P-Consciousness)

Percept-based or p-consciousness is based on the awareness or perception of one's sensations and hence all forms of life have, to a certain extent, some form of consciousness. Percept-based or p- consciousness is nothing more than the awareness of one's perceptions. The simplest bacteria are able to distinguish food from toxins and move towards the former and away from the latter. This is a very primitive form of p-consciousness. One may say that bacteria are conscious or aware of food and toxins and act accordingly. This form of consciousness is chemical in nature.

There are many levels of percept-based p-consciousness including chemical, tactile, pain, thermoperception, kinesthetic, magnetic, auditory, olfactory, gustatory, equilibriception and visual forms of consciousness and each form of p-consciousness corresponds to its respective sensory channel or capability. P-consciousness is about an organism perceiving, sensing or experiencing environmental signals from its umwelt [9] as well as internal signals.

Chemical forms of consciousness involve the pheromones of prokaryotes, slime molds, plants, and even with higher forms of life such as social insects like ants and some vertebrates. Pheromones can signal alarms, food possibilities, mating opportunities, etc. Another form of chemical signaling is quorum detection in bacteria, single-cell eukaryotes, fungi and social insects in which the density of conspecifics can affect gene expression for single-cell organisms. These organisms are conscious of the con-specifics in their colony.

Other levels of consciousness include the following:

- Tactile consciousness is where organisms with a primitive sensory apparatus are aware of touch or pain or heat.
- Auditory consciousness involves sound signals such as tones, growls, whines, barks, hoots, bird songs, mating calls, and for humans, verbal language as well as non-verbal prosody.
- Olfactory consciousness involves the smell of flowers for insects and birds and the smell of prey for raptors or the smell of a potential mate for a variety of animals.
- Gustatory consciousness involves the tasting mechanism of the tongue and epiglottis of many different animal species.
- Magnetoception consciousness involves the detection of the earth's magnetic field for navigation purposes by organisms including bacteria, insects, lobsters, stingrays, turtles, and birds.
- Visual consciousness involves the detection of visual signals.
- Thermoperception consciousness includes sensitivity to temperature, proprioception for a kinesthetic sense, nociception for pain and equilibrioception for balance experienced by a wide variety of animals including us humans.
- A level of perception and hence the consciousness of plants exists including their awareness of light, moisture, temperature, gravity, touch (in the case of carnivorous plants and mimosa pudica), and chemical signals from other plants.
- Given that percept-based or p-consciousness is characteristic of the full range of living organisms, one can conclude that the reaction to this form of consciousness is instinctive (private communication Alice Braga Gastaldo) as opposed to concept-based a-consciousness, which can be acted upon with rational forethought.

4. Percept-Based and Concept-Based Thinking

Because consciousness is a sentient phenomenon before addressing the nature of percept and concept-based consciousness, i.e. p- and a- consciousness, I will first review my study of the emergence of language in *The Extended Mind: The Emergence of Language, the Human Mind and Culture* [4], where I make a distinction between perceptual and conceptual thinking. Concept-based thinking is associated with verbal language and is unique to human beings. In my study, I suggested that words and concepts co-emerged with humans. It is proposed that the origins of speech and the human mind emerged simultaneously as the bifurcation from percept-based to concept-based mental activity. This transition was a response to the chaos associated with the information overload that resulted from the increased complexity in hominin life. This complexity arose in part with the hominin control of fire that led to hominins living in large groups instead of nuclear family units in order to share the many benefits of fire. Living in large groups required the coordination of many individuals and the necessity of planning. It is surmised that the hominin brain could no longer cope with the richness of life solely on the basis of its perceptual sensorium and as a result a new level of order emerged in the form of conceptualization and speech. Speech arose as a way to control information and was also used as a medium for communication. "Rather than regarding speech as vocalized thought one may just as well regard thought as silent speech [4] (p. 5)."

The mechanism that allowed the transition from percept-based to concept-based mental activity was the emergence of speech. The words of spoken language are the actual medium or mechanism by which concepts are expressed or represented (actually re-presented). Words are both metaphors and strange attractors uniting many perceptual experiences in terms of a single concept. For example, the word 'water' is a strange attractor for all of our percepts of water, namely the water we drink, cook with, wash with, falls as rain, and is found in rivers, ponds, lakes and oceans. Spoken language and abstract conceptual thinking emerged simultaneously at exactly the same point of time as a bifurcation from the concrete percept-based thinking of pre-lingual hominins to the abstract concept-based thinking of Homo sapiens. It is suggested that it might also have been the moment that Homo sapiens or full humans first emerged from their hominin ancestors.

Before humans acquired language, their brain was a percept processor. With language, the mind emerged as a conceptualization engine so that the mind is the product of the brain and verbal language (mind = brain + language). Because the mind is capable of conceptualization it allowed for planning and complex social organization, all of which aided human survival.

For non-human organism sentience is percept based and is a universal property of all living organisms from the simplest bacteria to our hominin ancestors. Human beings, on the other hand, have both percept-based and concept-based sentience. As consciousness and sentience are essentially the same as Eugene T. Gendlin [2] has asserted, I suggest that all living organism are conscious in that they possess p-consciousness or percept-based consciousness. This is a rather bold statement but it is based on the work of Terrence Deacon [10] in his book *Incomplete Nature*. The argument for the universality of p-consciousness follows from Deacon's notion that all living organisms are teleodynamic systems or selves that operate in their own self-interest. In order to act in their own self-interest a living organism has to be conscious, i.e. aware of its surrounding and its internal state so that it can take the appropriate action to insure it is operating in its own self-interest. An agent cannot realize its purpose if it is not conscious of those things that bear on its wellbeing. P-consciousness or percept-based consciousness is nothing more than a living organism being aware of or experiencing something happening within their environment or within their body.

A-consciousness or concept-based consciousness is strictly restricted to human beings because we are the only organism capable of verbal language and hence conceptualization. A-consciousness is therefore being aware of our thoughts and knowing what we know and hence is basically listening to ourselves silently talking to ourselves. So, for me the really hard problem of consciousness is how we developed verbal language so necessary for a-consciousness. Unless we translate our percepts or the products of our p-consciousness into concepts or silent language they are not accessible "for use in reasoning and rationally guiding speech and action" and as such will not be part of our a-consciousness.

Despite attempts to explain the origin of language, we still do not have a totally satisfactory description of this phenomenon and hence not one of a-consciousness either. Our various explanations for the origin of language and hence a-consciousness are "just-so" stories or abductions, including my own [4]. This is why I claim that a-consciousness is a harder problem to understand than p-consciousness.

5. Complexity and Emergence and the Transition from Percept-Based Thinking and Consciousness to Concept-Based Thinking and Consciousness

I suggested above that it was the complexity of living together in clans instead of nuclear family groupings that gave rise to the emergence of verbal language and concept-based thinking and consciousness. I chose the words complexity and emergence purposely because I believe that this transition can be explained in terms of complexity theory, which grew out of general systems theory. Systems theory is the notion that a system must be analyzed in terms of the system as a whole and not as the sum of the components making up the system. Complexity theory concerns itself with the various interactions and feedback loops of the components of the system with each other and with the system as a whole. In a complex system the behavior of the system and its components are the simultaneous result of the top-down causes of the system on its components, the bottom-up causes of the components on the supervenient system and the lateral causes of the components on each other.

This results in the supervenient system having properties that none of its components have. In simple terms, the system is more than the sum of its parts.

I would suggest that living in a complex community where cooperation was key to survival through the maintenance of the camp fire that there emerged out of the complexity of human interactions the ability to communicate and formulate thoughts in the abstract medium of verbal language and conceptual thinking. Just as spoken language allows a community to coordinate, organize and plan its activities, silent speech or thought or a-consciousness allows individuals to coordinate, organize and plan their activities. Given that speech is vocalized thought and thought is silent speech, speech and thought are basically the same phenomena. Therefore, a-consciousness or thought is nothing more than listening to one's silent speech. A-consciousness is in a certain sense being aware of one's awareness and that requires thought in the form of silent speech to achieve.

Since non-humans are not capable of verbal language, I therefore conclude that non-humans are unaware of their thought processes and therefore their level of consciousness cannot match that of humans. This is a non-scientific proposition since we have no way of knowing how non-humans think or what they think about. Some would claim they do not think but only operate by instinct but I believe there is enough evidence that indicates that some non-human animals do indeed think given their ability to solve certain problems. I do not believe, however, that they possess a-consciousness. This is purely speculative as our theory of mind only extends to other humans who we assume act and think the way we do. I cannot assume a non-human thinks the way I do.

6. Discussion and Conclusion

I find Block's distinction between p-consciousness and a-consciousness very useful. My qualification of his work is that while p-consciousness is universal for all forms of living organisms, a-consciousness is not. I have argued that only humans are capable of a-consciousness because their ability to use language allows them to conceptualize and hence 'reason and rationally guide action,' one of Block's criteria for a-consciousness. I am ready to concede, however, that perhaps some other forms of life might be capable of conceptualization such as non-human primates based on the observations and experiments with non-human primates solving problems and communicating using words. The cases of Washoe who learned to use 350 ASL signs [11] and Kanzi who learned language through the use of a keyboard lexigram [12] are prime example. On the other hand, I am quite certain that bacteria, single cell eukaryotes, fungi and plants are not capable of a-consciousness. The question becomes where does on draw the line between those organisms capable of a-consciousness and those that are not. This is a question for those that study non-human animal behavior.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Block, N. On a confusion about a function of consciousness. *Behav. Brain Sci.* **1995**, *18*, 227–287. [CrossRef]
2. Gendlin, E.T. Implicit precision. In *Knowing without Thinking: Mind, Action, Cognition and The Phenomenon of the Background*; Radman, Z., Ed.; de Gruyter: Amsterdam, the Netherlands, 1995; p. 148.
3. Schneider, S.; Velmans, M. Introduction. In *The Blackwell Companion to Consciousness*; Velmans, M., Schneider, S., Eds.; Wiley: New York, NY, USA, 2008.
4. Logan, R.K. *The Extended Mind: The Emergence of Language, The Human Mind and Culture*; University of Toronto Press: Toronto, Canada, 2007.
5. Brandom, R. *Making It Explicit: Reasoning, Representing, and Discursive Commitment*; Harvard University Press: Cambridge, MA, USA, 1994.
6. Brandom, R. *Articulating Reasons: An Introduction to Inferentialism*; Harvard University Press: Cambridge, MA, USA, 2000.

7. Davidson, D. *Thought and Talk. Inquiries into Truth and Interpretation*; Oxford University Press: Oxford, UK, 1975.
8. Dummett, M. *Seas of Language*; Oxford University Press: Oxford, UK, 1993.
9. Von Uexküll, J. *Theoretical Biology*; Harcourt, Brace & Co.: New York, NY, USA, 1926.
10. Deacon, T. *Incomplete Nature: How Mind Emerged from Matter*; Norton: New York, NY, USA, 2012.
11. Gardner, R.A.; Gardner, B.T. *The Structure of Learning from Sign Stimuli to Sign Language*; Lawrence Erlbaum Associates: Hillsdale, NJ, USA, 1998.
12. Savage-Rumbaugh, S.; Lewin, R. *Kanzi: The Ape at the Brink of the Human Mind*; Wiley: New York, NY, USA, 1994.

information

MDPI

Article

Thinking in Patterns and the Pattern of Human Thought as Contrasted with AI Data Processing

Robert K. Logan [1,*] and Marlie Tandoc [2]

[1] Department of Physics, University of Toronto, 60 St. George, Toronto, ON M5S 1A7, Canada

[2] Book and Media Studies, University of St. Michael's College, University of Toronto, 60 St. George, Toronto, ON M5S 1A7, Canada; marlie.tandoc@mail.utoronto.ca

* Correspondence: logan@physics.utoronto.ca

Received: 5 March 2018; Accepted: 5 April 2018; Published: 8 April 2018

Abstract: We propose that the ability of humans to identify and create patterns led to the unique aspects of human cognition and culture as a complex emergent dynamic system consisting of the following human traits: patterning, social organization beyond that of the nuclear family that emerged with the control of fire, rudimentary set theory or categorization and spoken language that co-emerged, the ability to deal with information overload, conceptualization, imagination, abductive reasoning, invention, art, religion, mathematics and science. These traits are interrelated as they all involve the ability to flexibly manipulate information from our environments via *pattern restructuring*. We argue that the human mind is the emergent product of a shift from external percept-based processing to a concept and language-based form of cognition based on patterning. In this article, we describe the evolution of human cognition and culture, describing the unique patterns of human thought and how we, humans, think in terms of patterns.

Keywords: patterns; patterning; cognition; set theory; language; information; abductive reasoning

1. Introduction

Humans both deal with information and engage in creative thinking, while computers only process data according to the instructions of their human programmers. We believe humans are capable of both recognizing and creating patterns, but computers are only capable of recognizing the type of patterns that they have been programmed to look for. We believe computers are deduction engines that are also capable of induction, as is the case when they succeeded at mastering chess or Go. They are not capable of abductive reasoning or creating a story, however, and hence there are limits to their creativity. Abductive reasoning is a form of logical inference using imagination, in which the simplest and most likely hypothesis is posited to explain observed phenomena (see Peirce [1]). The claims made in this opening paragraph are argued for in the lead paper of this special issue by Braga and Logan [2]. The purpose of this paper is to provide background on human cognition (and cognition in general) for the debate of human versus computer cognition.

In Braga and Logan [2], the authors claim that the technological Singularity, the idea that through AI, certain computers will be able to create a level of intelligence greater than that of humans, is not possible because computers are not capable of abductive reasoning, imagination and creativity. We will argue in this article that patterning is an essential feature of human cognition and is a product of abductive reasoning and imagination, which are features that computers are not capable of. This article therefore supports the claim of Braga and Logan [2] that the hypothesis of the eventual emergence of the Singularity is not correct.

We believe that ability to recognize and create patterns led to the unique aspects of human cognition and culture. This shift led to the co-emergence of inter-related traits, all characterized by their ability to manipulate and restructure patterns, *via mathematics, verbal language, imagination and abductive reasoning*. Since imagination is a form of abductive reasoning and abductive reasoning is a form of imagination, we will pair the two notions as imagination/abductive reasoning for the purpose of discussing patterning in this article. Our justification for the pairing of abduction and imagination is that one of the definitions of abduction is the process whereby one finds the simplest and most likely explanation of what one has observed and it is imagination that enables one to carry out this process.

We also believe that the environmental pressure of information overload is a key factor in our proposed dynamic cognitive system of human intelligence, which drives us to continuously create media and technologies to be able to not only efficiently process patterns, but also to reconfigure patterns in novel ways. What led to this huge cascade of complex ways of manipulating and restructuring patterns at the dawn of homo sapiens? We propose, as an abduction, that these complex forms of pattern-structuring, namely mathematical, linguistic and creative thinking, bootstrapped themselves into existence via the human ability to simultaneously hold multiple parallel representations in both mind (online) and memory (offline), allowing us, with a greater representational sandbox, to continuously structure and restructure patterns in countless ways. How patterns become instantiated in *culture* will also be explored.

As hinted at by our title, we believe that the ability to recognize and create patterns is the key to understanding the nature of human cognition and culture as well, as specifically, uniquely human traits, namely the control of fire, the ability to deal with information overload, patterning, primitive set theory, verbal language, imagination, abductive reasoning, invention, art, religion, mathematics and science, are all interrelated. We have tried to order the description of these factors that make us humans unique more or less in the order in which they emerged. We say more or less because some developed or emerged simultaneously and their interrelationships are non-linear and are part of emergent dynamics. The narrative we are about to develop is a story, an abductive reasoning or simply just a guess of how we humans came to be this super intelligent being that we are. This story is not represented as science, since there is no way to test the validity of our conjectures, but we believe that it might provide some insight into the relationship between the unique features of the human modus operandi.

Our story begins with the ability of the genus, Homo, to control fire, which led to a new social structure of large numbers of people living together and in turn led to information overload and the need to co-ordinate the activities of a large group of people. The solution to deal with the information overload was the emergence of verbal language, which required patterning in the form of set theory, in which words acting as concepts represented all of the percepts related to each of those words/concepts. Verbal language had a transforming effect and made possible imagination/abductive reasoning, which in turn gave rise to new technologies, mathematics, science, artistic expression and religion. The set of relationships that we have just outlined and presented are by no means obvious, but we hope to convince you in this article that these characteristics of human cognition and culture are interrelated. The narrative we have just presented is linear, but the relationships among these elements that make us humans unique is by no means a linear one. The evolution of human culture is a story of a complex adaptive system, in which emergent dynamics are always the driving force of this evolution. As we develop our narrative, some elements will appear twice, given the non-linear nature of our subject and the linear nature of our medium of the written word. To further support some of our claims, we will draw from various neuroscience designs that shed light on some of the pattern-processing and restructuring tools implemented in our brains.

The environmental pressure of our proposed dynamic system is information overload, which drives us to continuously create technologies and media to be able to not only efficiently process patterns, but also to reconfigure patterns in novel ways. What led to this huge cascade of complex ways of manipulating and restructuring patterns at the dawn of homo sapiens? We propose that these complex forms of pattern-structuring, namely mathematical, linguistic and creative thinking, bootstrapped themselves into existence via the human ability to simultaneously hold multiple parallel representations in both mind (online) and memory (offline), allowing us, with a greater representational sandbox, to continuously structure and restructure patterns in countless ways. How patterns become instantiated in *culture* will also be explored. We will examine math, language, artistic and scientific thinking and imagination/abductive reasoning, as tools by which we process and restructure information to best suit our goals.

2. Human Control of Fire and the Ensuing Information Overload It Created

Faced with information overload, we have no alternative but pattern-recognition—Marshall McLuhan ([3], p. 132).

The human control of fire changed the way our human ancestors lived together. Before learning how to control fire, humans lived in nuclear family units consisting of the mother, the father and their non-adult children. As the children grew into adulthood they went off and formed their own nuclear families and hunted and gathered on their own so as not to interfere with their parent's food gathering activities ([4], Chapter 3; [5], p. 60).

With the control of fire, nuclear families banded together to form clans of related nuclear families to take advantage of the many benefits that fire offered such as (i) warmth, (ii) protection from predators, (iii) tool sharpening, and (iv) cooking, which increased the number of plants that could be made edible, killed bacteria and helped to preserve raw foods, such as meat. Living together in clans gave rise to new, more complex and larger social structures that bred a form of information overload because of the increased complexities of social interactions and the need to organize the activities of many people to gather food and to maintain the campfire.

This information overload of interacting with many people and carrying out more sophisticated activities led to the need for better communications to better co-ordinate social transactions and co-operative activities, such as the sharing of the benefits of fire, the maintenance of the hearth, food sharing, and large scale coordinated hunting and foraging. Communication in this new environment became essential, and it is, therefore, surmised that this gave rise to a new preverbal proto-language of social interactions that emerged with the proto-semantics of social transactions, which included greetings, grooming, mating, food sharing, and other forms of co-operation appropriate for clan living. The proto-syntax of social organization or intelligence included the proper ordering or sequencing of these transactions in such a way as to promote social harmony and avoid interpersonal conflict, and, hence, contribute to the survival and development of hominid culture. It was in this environment that verbal language emerged with a still more complex level of organization.

The mechanism by which verbal language emerged was that words acting as concepts came to represent all of the percepts associated with that particular concept. Reflecting the transition from external to more internal pattern-processing, language emerged as the transition from percept- to concept-based thinking, according to the thesis developed in the book *The Extended Mind: The Emergence of Language, the Human Mind and Culture* [4]. A word acts as a concept that connects all of the percepts related to that word. The use of a word like "water", representing the concept of water, triggers instantaneously all of the mind's direct experiences and perceptions of water, such as the water we drink, the water we cook with, the water we wash with, the water that falls as rain or melts from snow and the water that is found in rivers, ponds, lakes, and oceans. The word "water" also brings to mind all the instances where the word "water" was used in any discourses in which that mind participated either as a speaker or a listener. The word "water", acting as a concept and an attractor, not only brings to mind all "water" transactions, but it also provides a name or a handle for the concept of water, which makes it easier to access memories of water and share them with others or make plans about the use of water. Words representing concepts speed up reaction time and, hence, confer a selection advantage for their users. At the same time, those languages and those words within a language which most easily capture memories enjoy a selection advantage over alternative languages and words, respectively. Before humans had verbal language, the brain was a percept processor, but with language the brain became a mind that was capable of conceptualization and the ability to plan. The mind is the brain plus language.

Several experimental studies have found that providing labels (i.e., a word) facilitates the learning of categories, i.e., seeking patterns ([6,7]). In fact, a word acting as a handle of a concept can even override more *perceptual* category learning [8], further suggesting that human cognition does represent a shift away from only bottom-up learning, towards more top-down, internal representations and the manipulative restructuring of our world. Such restructuring becomes easier to facilitate and more graspable when you have a handle.

The mechanism that actually led to verbal language seems to have involved the ability of humans to create **patterns** by distinguishing set of objects or activities that are similar and differentiating them from other objects or activities. It is in this sense that **mathematics**, in the form of set theory, emerged. Logan and Pruska-Oldenhof [9], in a book entitled *A Topology of Mind – Spiral Thought Patterns, the Hyperlinking of Text, Ideas and More*, developed the thesis that "the human mind is intrinsically verbal and mathematical and that **language** and **mathematical thinking** co-emerged at the dawn of the emergence of Homo sapiens." They posited that the origin of mathematics in mind occurred in the form of the classification or the grouping of like things into sets or groups and giving a name to that set or group, namely a word, acting as a concept, used to represent that set of percepts. It is the same skill that eventually gave rise to modern set theory and group theory and, as they claim, preceded the skill of counting or enumeration and actually laid the foundation for it. Therefore, linguistic and mathematical thinking both seem to play an important role in how we process and manipulate patterns.

Further support for the notion that the human brain is *intrinsically* mathematical, and that this mathematical feature is highly intertwined with our linguistic abilities, is observed through the phenomenon of *statistical learning*. Statistical learning is the ability to extract complex regularities and **patterns** from our environments over time [10]. Infants, for instance, are remarkably good statistical learners. An infant can effectively learn what phonemes occur together in a sentence over time by implicitly picking up on the transitional probabilities of sounds [11]. This remarkable ability to pick up on **patterns**, that is, how elements of experience relate, is argued as the reason for infant's and children's remarkable ability to learn language. Indeed, it is this *mathematical* quality of the mind that actually allows us to learn and implement complex linguistic structures or **patterns**. This further suggests that mathematical and linguistic thinking are tightly intertwined. Equally as remarkable is not only the brain's ability to pick up on the *transitional* probabilities described above (i.e., what features frequently co-occur together), but also to form even more complex mathematical representations in mind, such as forming *distributional* probabilities of categories, where through exposure to exemplars, the brain is able to integrate these experiences to create a "prototypical representation or distribution (ibid.)."

What pattern-processing mechanisms are involved in statistical learning? One account of statistical learning suggests a two-step model, whereby **patterns** must be *extracted* and then *integrated* [12] Again, this type of model parallels our thesis that pattern-processing is not computationally uniform, but a dual-process of pattern-processing, whereby one involves the ability to perceptually pull out patterns (pattern-recognition) and the other to integrate them into a broader picture (which we call pattern-restructuring). For instance, when we are learning a category, this involves (1) noticing what features co-occur together (in a statistically reliable way), and (2) *integrating* them to form a more generalized group and an integrated and abstracted representation of the category (for example, a prototypical exemplar of what is a "cat"). This process of integrating and restructuring is essentially equivalent to creating "new information". In sum, we have argued that the mathematical properties of the mind and language are tightly intertwined. By feeding off each other, they likely played a role in the very fast, snowballing cognition of homo sapiens, with mathematical thinking pushing us to create sets and patterns, and linguistic thinking making it easier for us to integrate or bundle things together into these sets (i.e., a word acting as an anchor or attractor for multiple percepts). Mathematical thinking, then, acts as a tool to manipulate information into patterns, and language acts as the handle, making them easier to grasp.

Thus, we agree with McLuhan that pattern-recognition is, in fact, how we successfully deal with information overload, and pattern-restructuring is how we can use large amounts of information for our advantage. For instance, evidence that we are able to manipulate patterns to create categories that best suit our needs is seen in cross-cultural differences in categorization. An additional argument to more strongly support the claim that there are variations within sets, based upon the pressures of the environment, was pointed out by Unsworth et al. [13], in regard to the cross-cultural differences in the categorization of butterflies. In reference to past anthropological studies, these authors pointed out that in the Fore culture, there are very well-defined, rigid, and discriminated *sub-categories* of birds because birds held hunting value as food and, therefore, it was important to make very discrete sets about them. Diamond [14], however, pointed out that this culture lacked any of these discrete, distinctive sub-groupings for butterflies, because butterflies, unlike birds, held little tangible value to this culture. Unsworth et al. [13] then went on to compare this culture to the Tzeltal culture who also did not have discriminate categories for butterflies, but did for *butterfly larvae*, which they used as a food source and encountered as a threat to crop growth [15]. In other words, the value and role of an organism to a society changes how categories are created or how we want to structure and differentiate information. Thus, this anthropological example is a fascinating example of how the patterns that we focus on to create categories may vary based on their impact or lack of impact on our culture.

As a result, the human mind requires the scaffolding to be able to accommodate the flexible nature in which we create sets. Pattern-processing, the ability to process incoming information, is not enough, but we require the ability to flexibly *use and manipulate* this information to best suit our goals and our understanding of our environment.

3. The Co-Emergence of Math and Language: The Shift from External Pattern-Recognition to Internal Pattern-Restructuring

What kind of mental faculties differentiate human cognition from that of animals? What cognitive qualities make us uniquely human? In answer to these sorts of questions, Charles Darwin humbly stated that "the difference in mind between man and the higher animals, great as it is, certainly is one of degree and not of kind." Do our pattern-processing abilities then only reflect a difference in degree and capacity? Explicitly drawing from this quote, Mattson [16] echoes Darwin's idea and draws from a vast range of animal research to argue that the human brain is a result of its *superior pattern-processing* abilities and not necessarily any difference in the kind of these processes. Indeed, several studies have found "scaffolding" or more proto-cognitive abilities in higher primates, that do seem to only differ in terms of a difference in degree.

However, while we agree with Mattson [16] that increasingly complex and superior pattern-processing is a crucial hallmark of the human mind, we argue, however, that there may be some cognitive qualities that cannot be scaled or reduced down to the level of higher primates, as they are the product of *emergence*, which, by definition, cannot be reduced down to its individual parts. In fact, reducing certain human properties of cognition down to the "same" or a difference in degree may overlook many of the important qualities that make us human. Echoing this view, Cobley [17] argues that the entire field of biosemiotics radically "insists that humans are separated from other organisms by a difference in kind and a difference in degree." Others who also study dynamic systems and the emergent properties of such systems also view kind and degree similarly, and not as oppositional processes, whereby emergence is the process by which "a difference in degree *becomes* a difference in kind [18]". When do we *draw the line* of when a difference in degree becomes so different that it appears more appropriately as a difference in kind? This transition has been described as a quality of emergence, whereby "a difference in degree becomes a difference in kind". While the difference in degree of our remarkable ability to process patterns is undeniable, the hallmark of the human mind does appear, then, to be one of a difference in kind, rather of degree.

Instead of arguing that human cognition is superior only because we are just really good at pattern-processing in general (i.e., a difference in degree), we believe there is value in further breaking down what constitutes pattern-processing and what kinds of patterning humans *particularly excel at*. We will look at (1) pattern-recognition, or the ability to *perceive* and *extract* patterns from our environments, as well as (2) pattern-restructuring, or the ability to *manipulate* these patterns internally to *create new* patterns. Pattern-recognition is how we cope in a world of information overload, and pattern-restructuring is how we transcend it and create new information. It is in the latter kind of pattern-processing that humans excel, that is, the ability to flexibly manipulate patterns to suit our goals. We also suggest that this ability led to several uniquely human phenomena such as science, religion, and art. We argue that human cognition, as seen with mathematical, linguistic, scientific and imaginary thinking, represents a shift towards more pattern-restructuring based cognition and, thus, the hallmark of the human mind reflects a difference in degree and of kind.

4. The Nature of Patterns

Patterns paradoxically both *unify* and *divide* our world. It is through this process of differentiation, of grouping things that are similar from those that are dissimilar, that we are able to meaningfully extract *information.* We will argue that a crucial hallmark of human pattern-processing is the ability to flexibly manipulate patterns and restructure them in novel ways to best suit our goals. First, we will explore the concepts of information and patterns, and how patterns give rise to information. Then, we will explore pattern-processing in the context of the human mind and in terms of mathematics, language, science, social science, the arts and imagination/abductive reasoning, for the creation of cognitive tools.

5. What is Information? Forming a Pattern is Equivalent to Creating Information

This leaves us with the question of: How was it that humans were able to create the patterns for grouping things into sets of objects or activities that possess similar properties? Marlie Tandoc, in an independent study course supervised by Robert K. Logan, built upon the idea of a topology of mind and mathematics in mind by introducing the notion of "a set theory of mind", according to which a category or a set is the most basic form of information. Categories or sets can vary by their content and the way in which the elements of the category or set relate to each other.

What forms a pattern? It is paradoxical that the similarity of the elements of a set creates a difference between the very elements of the set and all of the things not in the set. Creating or defining a set automatically creates another set, which we will call the anti-set, consisting of all things not in the original set. Since information is a difference that makes a difference, according to Gregory Bateson [19], p.428, creating a pattern is equivalent to **creating** information. Without similarities there are no differences, because once one sees similarities it causes one to consider things that are not similar and, hence, different. Difference is merely the absence of similarity, just as zero is the absence of a number and dark is the absence of light.

We would like to propose that difference arises as a natural emergent property of similarity. Each time we make a similarity judgment, we *automatically* draw a boundary and everything outside that boundary is different. Thus, information can be defined as the *process* of this emergence of differences, created by making similarity comparisons. The idea that information should be viewed not as a noun, but as the verb of informing [4], follows from this idea that information should be studied as a *process*. If information is a difference that makes a difference according to Bateson [19], then it can also be said that information is a non-similarity that makes a non-similarity.

Imagine having apples and bananas randomly spread out in front of you and drawing a circle around all the apples. While we may have the intent of drawing a circle to *enclose* whatever is inside (i.e., these items are all similar in color and shape), a circle also creates a boundary to everything *outside* of it, that is, all the bananas (the anti-set). Similarly, a judgment of similarity, thus, naturally creates a judgment of a difference, and the byproduct of this process is what we call information. Therefore, there can be "no difference without similarity" [20] and vice versa.

Many times, while we do make similarity comparisons, differences seem to quite literally "pop out at us." For instance, in a visual search process, items that are different, particularly that only vary in one dimension, such as color, seem to perceptually "pop out" at us. This automaticity and perceptual salience is the power of the information as a difference and as an emergent property of similarity.

Similarity is the fundamental pattern processing *tool* of human cognition and difference is what *emerges from its use.* For instance, one empirical study found that children were only able to learn by comparing within-category similarities, whereas adults were able to learn categories from both within-category similarities and between-category differences [21]. This suggests that similarity comparisons may develop before differences. "Similar" is easier to see and process than "not-similar" or "different", just as it is easier to conceive of positive integers than it is to conceive of zero and negative integers. The ancient Greeks and Romans were able to conceive of positive integers but not of zero or negative integers.

We can view similarity as the absence of difference and eliminate one of the terms from our analysis and use similar and non-similar or different and non-different. It is easier for us to process and recognize less differences than it is for us to see more differences.

Identifying similarities is a form of pattern recognition and pattern recognition, in turn, is a way of dealing with information overload, as pointed out by Marshall McLuhan [4], p. 132: "Faced with information overload, we have no alternative but pattern-recognition." Recognizing similarities to create a category or a set also leads to the construction of words, and words in themselves are another way of dealing with information overload. As noted above, the emergence of verbal **language** was motivated by the **information overload** of humans living in close quarters around the campfire.

Creating a category involves making a generalization, with the result that the greater the specificity of the category or the set, the less general or encompassing it is. But, on the other hand, the more general the category, the less specificity it possesses. The four categories of i. living organisms, ii. animals, iii. dogs, and iv. cocker spaniels increase in their level of specificity but decrease in their order of generality. In other words, the more specific the similarity, the smaller the set or the less general it is. The complementarity of generality and specificity of categories parallels the complementarity of position and momentum in the Heisenberg uncertainty principal in quantum mechanics, which states that the more you know of an atomic particle's momentum the less you know about its position and vice-versa. We call the complementarity of specificity and generality of categories the LT uncertainty principle of generalization and specificity. The greater the scope of the generalization of a category, the less its specificity and the greater its specificity the less its generalization. It is a trade-off because you cannot have both. More generalized categories allow us to make more comparisons across a wider range of experiences, at the cost of losing specificity, i.e., there are less similarities between the elements of the category. With more specific categories, more similarities exist between the members of the category but the number of members is less than is the case with a more generalized category.

Is there a *sweet spot*? Words for basic-level categories such as dog, tree, flower, or bird are found in nearly all languages and cultures and have shown strong perceptual, learning and memory biases for this level of category [22]. Basic categories, instead, may be the most effective *level* that allows for an informational *sweet spot* of generalization and specificity, and, thus, acts as an adaptive mechanism for information, which evolved over time and converged as the easiest to learn. Over cultural evolution, the level of generality and specificity in words and categories that have come to "stick" may, thus, represent the most effective level of generality versus specificity in our linguistic environments.

Nonetheless, certain environmental or task demands may push us towards creating more general or specific categories. Evidence for this is that information-processing is highly relativistic, and the benefit of basic level categories can just as easily vanish as they seem to appear. In a case where you are moments away from a tiger about to pounce, the most valuable piece of information may be something more general, such as *predator* or *prey*, rather than the exact species of the tiger. For example, while basic level categories (such as that of bird, tiger, or bear) typically show the greatest perceptual advantage in category learning tasks when time is limited (~30 milliseconds) and you have less time to make a decision, generality wins over specificity in that you "spot the animal before you spot the bird [23]." These findings support that when time is short, you might see a predator before you see a Siberian Tiger. However, in cases when you have *more time* to make a decision, a higher level of specificity may be of more benefit. Vervet monkeys, for instance, have different signals for different types of predators, which allows them to take the best course of action for that specific type of predator. In summary, in this example, *time to make a response and context* play a role in which categories one tends to rely on. This further supports the notion that decision-making based on categories is highly dependent on context, in this case the time to make a response, underscoring the *relativistic nature of information*.

Tradeoffs between the general and the specific have become a huge area of interest for neuroscientists in complimentary learning systems theory. The two complimentary learning systems include (1) the ability to *integrate* information across different experiences, and (2) the ability to code for episodic specificity of individual experiences [24].

It makes sense that having a good sense of what a "chair" is (similarities), is necessary information to know how a "chair differs from a table" (differences). Similarly, facing a new and uncertain world, similarity knowledge may be more *flexible* in its capacity for novel comparisons. Only memorizing an optimized rule of how a chair is different than a table may not be that helpful when later having to differentiate between a chair and a couch. Instead, knowledge of the internal properties of the features likely to co-occur across chairs may be much more helpful, at least at first. Indeed, it may be advantageous to *initially* bias the system towards internal similarities, as this, in turn, can allow us to later use these similarities to *bootstrap* ([18,25]) diagnostic differences into existence and to do so across a wide range of contexts.

In summary, we have argued that biasing attention towards similarities, as the crux of pattern-processing, is beneficial, as it (1) constrains the meaningful differences to emerge (as information), and (2) allows more *flexible* use of this information for *novel* comparisons.

6. Letting Go: Pattern-Restructuring as the Mechanism of Novel Ideas, Language and Imagination

"The **spoken word** was the first technology by which man was able to **let go** of his environment in order to **grasp it in a new way**."—Marshall McLuhan.

Language, imagination, abductive reasoning, invention (as in technology), art, religion and science are unique characteristics of humans that no other animals possess and seem to be related to each other, as we will claim, and this idea is supported by a number of other scholars that will be referenced in this section.

Marshall McLuhan and Robert K. Logan [26] wrote:

> *If one must choose the one dominant factor which separates man from the rest of the animal kingdom, it would undoubtedly be language. The ancients said: 'Speech is the difference of man' . . . It is the medium of both thought and perception as well as communication.*

A number of authors have made a connection between language and imagination. Among these are:

Daniel Dor [27], author of the book, "The Instruction of Imagination: Language as a Social Communication Technology", characterizes "language as a functionally specific communication technology, dedicated to the instruction of imagination: with language, and only with it, speakers can make others imagine things without presenting them with any perceptual material for experiencing."

Eric Reuland, author of the articles, "Imagination, planning, and working memory: the emergence of language [28]**"** and "Language and imagination: Evolutionary explorations [29]", shows the intimate relationship between language, imagination and planning. He argues that language makes imagination possible [28] where he explores the relation between imagination, planning and language. He [29] also claims that imagination is "*the language lab*, producing both science and poetry".

Paul Crowther [30]), author of "Imagination, language, and the perceptual world: a post-analytic phenomenology", argues "that language directs imagination, empirically speaking in adult life, but that the ontogenesis of language presupposes the role of imagination (ibid., 40)." He claims that "as imagination is a mode of thought, . . . it follows that there must be some key relation to language, also, insofar as thought in its fullest sense, is centered on language. The relation between imagination and language is, in fact, a vital one, empirically speaking. It dominates how imagination is exercised (ibid., 43)."

Mark Mattson [16], author of the article, "Superior pattern processing is the essence of the evolved human brain", begins his review article in *Frontiers in Neuroscience* with the following provocative remark in the first two sentences of his abstract:

> *Humans have long pondered the nature of their mind/brain and, particularly why its capacities for reasoning, communication and abstract thought are far superior to other species, including closely related anthropoids. This article considers superior pattern processing (SPP) as the fundamental basis of most, if not all, unique features of the human brain including intelligence, language, imagination, invention, and the belief in imaginary entities such as ghosts and gods.*

What stopped us in our tracks was Mattson's pairing of "intelligence, language, imagination, invention" with "the belief in imaginary entities such as ghosts and gods". As practitioners and students of science, we associated "intelligence, language, imagination/abductive reasoning, and invention" with science, engineering and the arts. On the other hand, we associated "belief in imaginary entities such as ghosts and gods" with faith, at best, and superstition, at worst. We were intrigued by Mattson's abstract and so we read the whole article. As we discussed its thesis, it suddenly occurred to us that perhaps abductive logic or thinking was the link between these two seeming disparate categories of "intelligence, language, imagination and invention", on the one hand, and the belief in "imaginary entities such as ghosts and gods", on the other hand.

Abductive logic is unlike deductive logic and inductive logic. Let us explain. With deductive logic, you begin with two axioms that you believe to be self-evident and deduce a conclusion. Socrates is a man. All men are mortal. Therefore, Socrates is mortal. With inductive logic, you list all examples where your conclusion is true and you assume or guess, therefore, it is always true. Socrates was mortal. Aristotle was mortal. Newton was mortal. Einstein was mortal. Therefore, all men are mortal. With abductive reasoning, one observes a set of data and one then guesses or *imagines* or invents a hypothesis that explains that data, that is the simplest and most likely. All three forms of logic involve guessing in one way or another. With deductive logic, one makes the guess that one's starting axioms are correct. With inductive logic, one makes the guess that if the statement is true for all the examples that one is able to compile, then it must always be true for all possible cases.

Each of these three forms of logic have different applications:

Mathematics, for the most part, proceeds by way of deductive logic. The axioms that parallel lines (i) remains the same distance apart; (ii) converge; or (iii) diverge, give rise to three forms of geometry, respectively: (i) plane; (ii) Riemannian; and (iii) Lobachevskian.

Inductive reasoning is used for forecasting and cannot guarantee the truth of its conclusion, but only suggest that it is most likely. In most cases, the greater the sample size, the more reliable is the conclusion, but this assertion itself is another example of inductive reasoning.

Abductive thinking, as used in science, only asserts that its conclusions are the simplest and most likely, but must be subjected to constant testing. The criteria for a conclusion to be considered as a scientifically valid conclusion is that it must be falsifiable, as suggested by Karl Popper [31].

So, what is the connection between "intelligence, language, imagination, invention" and "the belief in imaginary entities such as ghosts and gods"? Belief in imaginary entities, such as ghosts and gods, parallels abductive thinking and requires imagination. Belief in, or even suggesting a hypothesis that explains observed phenomena in science, also involves abductive reasoning and also requires imagination.

Imagination is a key requirement in both science and magical thinking, both of which create or recreate a reality for the one imagining the science hypothesis or the magical thinking. José Monserrat Neto [32] concurred with this when he suggested "that the capacity of imagination is essential to understand the creative way in which human beings learn and (re)construct their reality."

A theory or hypothesis in science is an imaginary entity which entails the belief that it might be actual or true. Science and the belief in ghosts or gods is, therefore, parallel to a certain degree. They diverge in that the scientist accepts their hypothesis might be false and hence needs to be constantly tested empirically, whereas the religious or ghost believer does not accept the notion that their belief may be false, but accepts the validity of their belief on faith alone without a need to test their hypothesis.

The human ability to create these hypotheses to account for data involves imagination/abductive reasoning and the ability to combine seemingly disparate patterns and restructure them in new ways. Although a deterministic, machine-like system (artificial intelligence, for instance) is able to use both deductive logic (if-then statements) and inductive logic (machine learning), the use of effective *abductive* reasoning may be a uniquely human phenomenon. Abductive reasoning is the result of a mathematical mind with the ability to create novel categories or patterns between otherwise seemingly disparate elements via *pattern restructuring*.

The three different forms of thought or logic, deductive, inductive and abductive, can be seen as ways in which we can navigate information landscapes amongst the different levels of specificity and generality of our concepts that we defined above in the LT uncertainty principle of generalization and specificity (see Figure 1). Deductive logic involves moving from the general to the specific, and, inductive, from the specific to the general. Both of these act as a one-way street of logic. Abductive thought requires a more horizontal, domain-general movement that focuses on forging *new connections* across ideas and consideration of context, which we believe is uniquely human.

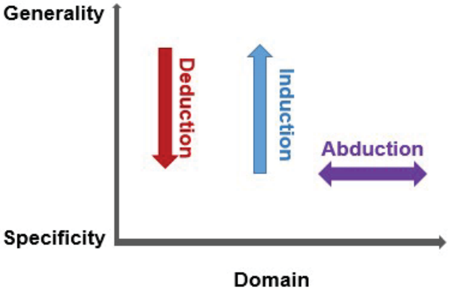


Figure 1. Generality versus Specificity for Deduction, Induction and Abduction.

Further supporting the importance of having vivid, detailed, episodic experiences, is that this system may be how we also imagine *future* events. *That is, episodic memory is the foundational, sensational landscape we use to feel like we are there when imagining future events, and also involves the ability to restructure patterns on a whim to create internal worlds*.

The same storage of multiple representations via pattern separation and holding multiple representations in mind via working memory, and having access to episodes, are the processes that seem to allow us to *imagine* future events.

7. Pattern Restructuring: Abductive Reasoning and Creativity

We argued above that abductive reasoning may be able to explain both scientific and superstitious thinking, but where they differ is in terms of falsifiability. Abductive reasoning that leads to such a vast range of what makes us human stretches from science and intelligence, to religious faith and belief in ghosts [16]. Abductive thinking, therefore, seems to be a uniquely human phenomenon. Abductive thinking involves being able to restructure connections between ideas, to create novel ideas to bridge often seemingly distant ideas and to create hypotheses that can explain and unify observations, and to create "just so stories" and coherent narratives. It requires *creativity* and the ability to connect and restructure patterns

How do we get something seemingly novel from previous information that seems disconnected? Is creativity a magical "and then there was light" phenomenon and, seemingly, something emerging from nothing? Not necessarily. What if we just create new connections between previously existing ideas? This would explain how something seemingly novel can arise from pre-existing, disconnected bits of information. In terms of our "set theory" of mind, creativity comes from this ability to forge *novel connections* between sets, even sets that may have little overlap at first glance. In other words, the ideas and content remain constant, but it is the *connections between* these patterns that give way to creativity, imagination and abductive reasoning, which then results in new ideas and new content. The following definition of creative insight is particularly helpful explaining this idea:

> *(Creative) insight seems to involve (1) an existing state of mind or set of mental structures relevant to the topic and (2) a moment of realization, consequent to new information or a sudden new way of looking at old information, resulting in (3) a quick restructuring of the mental model, which is subjectively perceived as providing a new understanding [33].*

Csikszentmihalyi's definition of creative thought implies that we have to have these pre-existing sets, formed via our pattern-processing tools, and then creativity/**pattern-creation** emerges as pattern-restructuring and bridging novel connections between these already existing ideas. This helps explain how novel, creative ideas can come about from seemingly out of nowhere (i.e., the "ah-ha" moment), though not necessarily through any new content, but simply different connections between existing content, which, in turn, creates new content though the new connections. Note that creative thought or abductive reasoning is not finding something from nothing, it is the ability to find something from a group of initially seemingly unrelated things. It is creating new sets of syntactical structures, which, as a result, creates new meaning and hence new knowledge. Everything was there to begin with, we just were able to realize it, or rather, in McLuhan's words, let go of it to grasp it in a new way, and this process in itself gives rise to new meaning and new information.

Indeed, just as we argued that math and language are ultimately boiled down to pattern restructuring tools, creativity, also, is the ability to restructure information to solve problems or create *novel* ideas. Again, recall Czkimenthalyi's first step of creativity, which involves "an existing state of mind or set of mental structures ... ", and further supports our view that vivid episodic experiences are required to lend themselves to the increasingly complex pattern restructuring of these ideas, as without this raw resource to return to, we become **fixated** only on one idea or pattern.

Our ability to imagine and create allows us to create increasingly complex ideas, from technology, art, science, and religion. Ultimately then, many of the qualities that make us *human* stem as an emergent property of our ability to *restructure patterns* to zap novel ideas into existence. But of course, all the material is there to begin with, the human mind is just able to recognize, extract, integrate, synthesize, restructure, and *realize it* into existence.

8. No Signal, Without Noise: The Role of "Randomness" in Pattern-Restructuring

As argued above, creativity involves being able to take seemingly unrelated ideas or sets and restructuring them to make novel ideas. For instance, even at the neuronal network level, the brain seems to have a hard-wired mechanism that works *to increase our chances* of the finding of novel ideas, by pushing us to restructure our patterns in novel ways. Connections of neurons, which could represent an idea, for instance (i.e., a consistent pattern of neurons fire every time you think of a "dog"), have been found to fire in unpredictable, random connections, called *stochastic resonance*. While it is still debated, the reason for these "noisy" neurons, it has been recently suggested, can be seen as the brain *testing out new connections*. This testing out of new neuronal combinations is argued to prevent decisional deadlock, to let us make mistakes to learn from, and might even act as a basis for *creativity* (see Deco et al. [34]). This final point on stochastic resonance as the grounds for creativity, is quite the assertion, but follows the idea that certain neuronal computations can be hierarchically abstracted [35] to assist in explaining increasingly complicated, abstracted ideas.

Randomness and chance indeed play a role in creativity, and our brains may take advantage of this randomness to push towards bridging otherwise disparate ideas. The role of "randomness" and chance is analogous to incredibly creative design we observe in evolution and natural selection. A random mutation that made a giraffe's neck slightly longer was the necessary random push needed to lead to the impressive design of the giraffe's famously elongated neck. Similarly, neuronal stochastic resonance, essentially a mutation of neural firing patterns, may be the initial "nudge" needed to lead to a cascade of pattern-restructuring processes, lending itself to the restructuring of ideas, to lead to an impressively novel idea, or creativity. All biological mutations are not necessarily helpful, but we require this variation, and over time and chance, one mutation may prove to be an adaptation. Random neuronal firing may similarly prove to be a creative adaptation, as over time some might push the mind towards finding otherwise unseen connections that turn out to be very useful.

Thus, just as natural selection requires this bit of randomness and open-endedness, the human mind is also remarkably noisy and open-ended. The human brain has more than 100 trillion connections between its neurons. There is a hugely vast number of connections we can forge and reforge to make new ideas. In his new book, Daniel Dennet [36] wrote that "evolution is all about turning 'bugs' into 'features,' turning 'noise' into 'signal,' and the fuzzy boundaries between these categories are not optional; the opportunistic open-endedness of natural selection depends on them." Analogous to this, we would argue that the flexible human mind also depends on such opportunistic open-endedness and the ability to restructure patterns to allow for creativity and imagination/abductive reasoning.

One person's noise might be another person's information. "The user is the content", as McLuhan once opined [37] (p. 51). Terrence Deacon [38] point out that "noise can be signal to a repairman." As humans are always looking for unseen connections and **patterns** to better understand our world, we are all repairmen in a sense. As repairmen, we have the appropriate tools to extract signal out of what may have initially appeared as noise. We, as humans, have the cognitive tools necessary to create information.

9. Brand New: Thinking in Patterns

To further support our hypothesis that patterning is a critical part of human cognition, we describe the role of patterning in a number of human intellectual activities, ranging from the fine arts, music and verbal language to mathematics, natural science and the social sciences.

Verbal Language: We have already suggested that the emergence of words as concepts that describe all the percepts associated with that concept is a form of patterning. What applies for semantics also applies to syntax, as grammar also represents a form of patterning. Pragmatics, the third aspect of understanding verbal language, underscores the importance of the pattern of pragmatics in communicating meaning.

Mathematics: Mathematics entails patterns in a multitude of ways. A sequence, for example, is a string of objects, like integers, that follow a particular pattern, such as the set of all even numbers or the set of all integers that are divisible by the integer, n. The Fibonacci sequence is another example of a sequenced pattern. The elements in the sets of all triangles, or of all rectangles, etc., each share a common pattern. Patterns abound in algebra and geometry in numbers too great to describe in this article.

Music: Patterns abound in all forms of music, from the simplest folk song to the most complex symphony, concerto or sonata. A melody is a particular pattern of notes. Rhythmic patterns are an essential form of music. Many musical dance forms have a particular rhythmic pattern, such as the waltz, the mambo, the tango or the samba. The fugue, the sonata form, the rondo form, the round, and counterpoint are examples of the many forms or patterns of music.

The Visual Arts and Architecture: Patterns are an essential part of both the fine and decorative arts, as well as architecture. Symmetry is used throughout the visual arts and architecture and is one of the elements of what we consider to be beautiful, not just in the arts but also what we consider as human beauty. Beginning in the Renaissance, the pattern of perspective, with its vanishing point, is an essential feature of the visual arts.

Engineering: Patterns are an essential part of successful engineering projects, particularly in mechanical and civil engineering.

Science and Social Science: Detecting patterns is an essential part of the natural sciences. The Copernican Revolution, Kepler's Laws of Planetary Motion, Newton's Three Laws of Motion and the Law of Gravity, and Maxwell's Equation of Electromagnetism, all involve identifying patterns in nature. The discovery of the Periodic Table of Elements in chemistry entailed finding the patterns of chemical bonding. The classification of biological species and genera is another example of the uncovering of a pattern in the world of biology. Meteorology is based on the identification of weather patterns. The law of supply and demand is an example of a pattern in economics. Patterns play a role in other social sciences such as sociology, anthropology, law, psychology and cognitive science. Patterning is an essential tool in all of the sciences and the social sciences. In some sense, patterns are a way to understand the order we find in nature and the social interactions of humans.

10. Patterns in Human Culture

We are not crediting all of our creative thought, that is, our ability to form ideas and concepts, to *randomness.* Perhaps randomness is an offline way in which we can prevent creative deadlock that increases our flexibility. Another way in which we may connect ideas is through pattern-completion in auto-association networks, as suggested by Hebb [39]. Pattern completion occurs when there is high overlap between two experiences, and the brain encodes this information with higher neural similarity in an integrated representation. This may be another mechanism that can help us make new connections.

We are not crediting all of our creative thought to *randomness*, or our ability to further *unravel* and form ideas and concepts, but it is perhaps an offline way in which we can prevent creative deadlock and increase our flexibility, within our various ways of restructuring patterns. Similarly, Harnad [40], in his paper titled "Creativity: Method or Magic", argued for Pasteur's Dictum, whereby "chance favours the prepared mind", for the case of creativity. In our thesis, having access to strong representations of episodic memories and percepts is a way of "preparation" for the creative patterns that can emerge from it.

Although we have focused on the more concrete ways in which we form and manipulate patterns, as seen with the role of randomness, the mind is largely probabilistic and we have enough variation in how we form patterns. As a result, patterns have variation and, thus, become even more powerful, as they can adapt to different contexts.

However, what happens to these patterns as they form? How do they become instantiated in culture? The topology of the mind is structured by the categories that the mind creates. The categories of an individual's mind are subject to natural selection. Those categorizations that increase the fitness of those that possess them or the culture in which they thrive tend to survive.

While we have focused primarily on the *process* of pattern restructuring, that is, how the mind forges patterns, now we will also briefly focus on the *products* of these processes, that is, how properties of the mind itself become instantiated in these patterns over time, rendering them easier to learn.

Each time information is transmitted, whether through neural synapses or verbal speech, it becomes susceptible to change. As a result, the pattern boundaries (the meaningful differences) that come to stick over time are often the easiest to learn, that is, the ones that make it most frequently through the *mind's* cognitive filter.

In this story, the mind, as the medium, then becomes the message (McLuhan). Indeed, this system, in which the most "fit" patterns come to stick and with information constantly driving us to seek new ways to restructure, is the dynamic system. This iterative process is the ultimate driving force of patterns, across time and space.

Dawkin's [41] meme theory is perhaps the most famous, and controversial, instance of this type of transmission of patterns (or memes), whereby patterns come to best "fit" in the human mind. In Dawkin's view, memes (or patterns) act as viruses that "infect" the mind. We disagree, as we view pattern-restructuring as the "kind" of pattern-processing that humans excel at. As a result, we place the human mind in a much more *active* role than Dawkins poses. Indeed, our view falls nicely in line with many critics of Dawkin's meme theory, because the mind is *creating* and recreating patterns, it is an active agent that *creates* novel ideas and not just an "imitator" of ideas [42]. Our central tenant of our thesis is that we do not just *recognize* patterns, but we internalize them, then we manipulate and restructure them to suit our goals. The mind is not just an imitator or pattern processor, it is an inference creator that actively restructures its internal representations. It is this second aspect that is more characteristically human, that we are able to flexibly restructure our worlds. Our thesis is one that supports the human mind as an active agent, not just a helpless victim of its environment.

11. Conclusions

Mathematical, linguistic, creative, imaginary, and abductive thinking are the modes by which our mind restructures patterns. We have described how patterns, formed via recognition and restructuring, become instantiated in culture. To conclude, we would suggest patterning, particularly the ability to create or see new unexpected patterns, is the key to creativity in the arts, science and religion, three domains of spirituality that might have more in common than is generally recognized.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Peirce, C.S. *Collected Papers*; Harvard University Press: Cambridge, MA, USA, 2012; pp. 1931–1958.
2. Braga, A.; Logan, R.K. The emperor of strong AI has no clothes: Limits to artificial intelligence. *Information* **2017**, *8*, 156. [CrossRef]
3. McLuhan, M. *Counterblast*; McClelland and Stewart: Toronto, ON, Canada, 1969.
4. Logan, R.K. *The Extended Mind: The Emergence of Language, the Human Mind and Culture*; University of Toronto Press: Toronto, ON, Canada, 2007.
5. Donald, M. Mimesis and the executive suite. In *Approaches to the Evolution of Language*; James, H., Michael, S.-K., Chris, K., Eds.; Cambridge University Press: Cambridge, UK, 1998; pp. 44–67.
6. Fulkerson, A.L.; Waxman, S.R. Words (but not tones) facilitate object categorization: Evidence from 6-and 12-month-olds. *Cognition* **2007**, *105*, 218–228. [CrossRef] [PubMed]
7. Waxman, S.R.; Markow, D.B. Words as invitations to form categories: Evidence from 12-to 13-month-old infants. *Cogn. Psychol.* **1995**, *29*, 257–302. [CrossRef] [PubMed]
8. Plunkett, K.; Hu, J.F.; Cohen, L.B. Labels can override perceptual categories in early infancy. *Cognition* **2008**, *106*, 665–681. [CrossRef] [PubMed]
9. Logan, R.K.; Izabella, P.-O. *A Topology of Mind—Spiral Thought Patterns, the Hyperlinking of Text, Ideas and More*; Springer: New York, NY, USA, 2018; in press.
10. Fiser, J.; Aslin, R.N. Statistical learning of higher-order temporal structure from visual shape sequences. *J. Exp. Psychol. Learn. Mem. Cogn.* **2002**, *28*, 458. [CrossRef] [PubMed]
11. Saffran, J.R.; Aslin, R.N.; Newport, E.L. Statistical learning by 8-month-old infants. *Science* **1996**, *274*, 1926–1928. [CrossRef] [PubMed]
12. Thiessen, E.D.; Kronstein, A.T.; Hufnagle, D.G. The extraction and integration framework: A two-process account of statistical learning. *Psychol. Bull.* **2013**, *139*, 792. [CrossRef] [PubMed]
13. Unsworth, S.; Sears, C.R.; Pexma, P.M. Cultural influences of categorization processes. *J. Cross-Cult. Psychol.* **2005**, *36*, 662–688. [CrossRef]
14. Diamond, J. Zoological Classification of a Primitive people. *Science* **1966**, *51*, 1102–1104. [CrossRef] [PubMed]
15. Hunn, E. The Utilitarian factor in folk biological classification. *Am. Anthropol.* **1982**, *84*, 830–847. [CrossRef]

16. Mattson, M. Superior pattern processing is the essence of the evolved human brain. *Front Neurosci.* **2014**, *8*, 265. [CrossRef] [PubMed]
17. Cobley, P. Difference in Kind or Difference of Degree? In *Cultural Implications of Biosemiotics*; Springer: Amsterdam, The Netherlands, 2016; pp. 29–44.
18. Piers, C. Emergence: When a difference in degree becomes a difference in kind. In *Self-Organizing Complexity in Psychological Systems*; Craig, P., John, P.M., Joseph, B., Eds.; Rowman and Littlefield: Lanham, MD, USA, 2007; pp. 83–110.
19. Bateson, G. *Steps to an Ecology of Mind*; University of Chicago Press: Chicago, IL, USA, 1973.
20. Gentner, D.; Markman, A.B. Structural alignment in comparison: No difference without similarity. *Psychol. Sci.* **1994**, *5*, 152–158. [CrossRef]
21. Hammer, R.; Gil, D.; Daphne, W.; Shaul, H. The development of category learning strategies: What makes the difference? *Cognition* **2009**, *112*, 105–119. [CrossRef] [PubMed]
22. Rosch, E.; Mervis, C.B.; Gray, W.D.; Johnson, D.M.; Boyes-Braem, P. Basic objects in natural categories. *Cogn. Psychol.* **1976**, *8*, 382–439. [CrossRef]
23. Macé, M.J.M.; Joubert, O.R.; Nespoulous, J.L.; Fabre-Thorpe, M. The time-course of visual categorizations: You spot the animal faster than the bird. *PLoS ONE* **2009**, *4*, e5927. [CrossRef] [PubMed]
24. O'Reilly, R.C.; Norman, K.A. Hippocampal and neocortical contributions to memory: Advances in the complementary learning systems framework. *Trends Cogn. Sci.* **2002**, *6*, 505–510. [CrossRef]
25. Carvalho, P.F.; Goldstone, R.L. The benefits of interleaved and blocked study: Different tasks benefit from different schedules of study. *Psychon. Bull. Rev.* **2015**, *22*, 281–288. [CrossRef] [PubMed]
26. McLuhan, M.; Robert, K.L. Alphabet, Mother of Invention. *ETC* **1977**, *34*, 373–382.
27. Dor, D. *The Instruction of Imagination: Language as a Social Communication Technology*; Oxford University Press: Oxford, UK, 2015.
28. Reuland, E. Imagination, planning, and working memory: The emergence of language. *Curr. Anthropol.* **2010**, *51* (Suppl. S1), 99–110. [CrossRef]
29. Reuland, E. Language and imagination: Evolutionary explorations. *Neurosci. Biobehav. Rev.* **2016**. [CrossRef] [PubMed]
30. Crowther, P. Imagination, language, and the perceptual world: A post-analytic phenomenology. *Cont. Philos. Rev.* **2013**, *46*, 37–56. [CrossRef]
31. Popper, K. *The Logic of Scientific Discovery*; Routledge: London, UK, 1959. (In German)
32. Csikszentmihalyi, M. Society, culture, and person: A systems view of creativity. In *The Systems Model of Creativity*; The Collected Works of Mihaly Csikszentmihalyi; Springer: Amsterdam, The Netherlands, 2014; pp. 47–61.
33. Monserrat Neto, J. The Wings of Imagination—The missing link in the origin of consciousness? *Neurociências* **2014**. Available online: http://www.academia.edu/1515781/The_Wings_of_Imagination_-_the_missing_link_in_the_origin_of_consciousness (accessed on 7 April 2018).
34. Deco, G.; Rolls, E.T.; Romo, R. Stochastic dynamics as a principle of brain function. *Prog. Neurobiol.* **2009**, *88*, 1–16. [CrossRef] [PubMed]
35. Ballard, D.H. *Brain Computation as Hierarchical Abstraction*; MIT Press: Cambridge, MA, USA, 2015.
36. Dennett, D.C. *From Bacteria to Bach and Back: The Evolution of Minds*; WW Norton & Company: New York, NY, USA, 2017.
37. Logan, R.K. *McLuhan Misunderstood: Setting the Record Straight*; The Key Publishing House: Toronto, ON, Canada, 2013.
38. Deacon, T.W. *Incomplete Nature: How Mind Emerged from Matter*; WW Norton & Company: New York, NY, USA, 2011.
39. Hebb, D.O. *The Organization of Behavior*; Wiley & Sons: New York, NY, USA, 1949.
40. Harnad, S. Creativity: Method or magic? *Hung. Stud.* **2006**, *20*, 163–177. [CrossRef]
41. Dawkins, R. *The Selfish Gene*; Oxford University Press: Oxford, UK, 2016.
42. Atran, S. The trouble with memes. *Hum. Nat.* **2001**, *12*, 351–381. [CrossRef] [PubMed]

information

Commentary

Love, Emotion and the Singularity

Brett Lunceford

Independent Researcher, San Jose, CA 95136, USA; brettlunceford@gmail.com; Tel.: +1-408-816-0834

Received: 31 July 2018; Accepted: 31 August 2018; Published: 3 September 2018

Abstract: Proponents of the singularity hypothesis have argued that there will come a point at which machines will overtake us not only in intelligence but that machines will also have emotional capabilities. However, human cognition is not something that takes place only in the brain; one cannot conceive of human cognition without embodiment. This essay considers the emotional nature of cognition by exploring the most human of emotions—romantic love. By examining the idea of love from an evolutionary and a physiological perspective, the author suggests that in order to account for the full range of human cognition, one must also account for the emotional aspects of cognition. The paper concludes that if there is to be a singularity that transcends human cognition, it must be embodied. As such, the singularity could not be completely non-organic; it must take place in the form of a cyborg, wedding the digital to the biological.

Keywords: cognition; cyborg; embodiment; emotion; evolution; love; singularity

1. Introduction

This essay takes up where Adriana Braga and Robert Logan [1] left off in their recent essay, "The Emperor of Strong AI Has No Clothes: Limits to Artificial Intelligence," in which they argue against the notion of the "singularity" or a point at which computers become more intelligent than humans. However, rather than focusing on intelligence, this essay extends Braga and Logan's discussion of emotion and focuses on cognition, exploring what it means to think and what makes human cognition special. I suggest that the foundation for this exceptionalism is emotion.

Cognition is a slippery thing and despite considerable study, we are far from fully understanding how humans think. The question of what it means to think, to be sentient, is one that has likely plagued humanity since we have been able to articulate the question. We have some hints of what it means to think from people like René Descartes [2] (p. 74), who proclaimed, "it is certain that this 'I'—that is to say, my soul, by virtue of which I am what I am—is entirely and truly distinct from my body and that it can be or exist without it." In other words, there is something in us that goes beyond biology, a kind of self-awareness that one exists. But the notion of sentience is a bit more complicated than that. As Clark [3] (p. 138) argues, "There is *no self*, if by self we mean some central cognitive essence that makes me who and what I am. In its place there is just the 'soft self': a rough-and-tumble, control sharing coalition of processes—some neural, some bodily, some technological—and an ongoing drive to tell a story, to paint a picture in which 'I' am the central player." We are more than the information processed in our brains, which complicates the posthumanist dream of having one's consciousness uploaded into a computer to live forever (unless someone pulls the plug, of course). As Hauskeller [4] (p. 199) explains, "the only thing that can be copied is information, and the self, *qua* self, is not information." In short, although we understand the processes of cognition (e.g., which segments of the brain are active during certain activities), we are far from understanding exactly how sentience emerges.

Even if we were to understand how sentience emerges in a human being, this still would not bring us any closer to understanding how sentience would emerge in a synthetic entity. Some have defined thinking machines tautologically; for example, Lin and colleagues [5] (p. 943) "define 'robot' as *an engineered machine that senses, thinks, and acts*." Although this is a convenient way to define thinking, it fails to get us any closer to understanding what it is that separates humans from synthetic entities in terms of cognition. As an aside, I would note that I use the term synthetic beings deliberately, because there is no reason why an entity in possession of artificial intelligence would necessarily have a body in the way that we imagine a robot to have. Of course, there would need to be some physical location for the entity to exist, as an artificial intelligence that we would create would require some form of power source and hardware, it need not be in one specific location and could be distributed among many different machines and networks.

My point in all of this is that we tend to take an anthropocentric view of robots and then measure them up against how well they mimic us. After all, the Turing Test measures not intelligence but rather how well they can deceive us by acting like us, when it is quite possible that they may actually engage in a kind of thinking that is completely foreign to us [6]. As Gunkel [7] (p. 175) explains,

> There is, in fact, no machine that can "think" the same way the human entity thinks and all attempts to get machines to simulate the activity of human thought processes, no matter what level of abstraction is utilized, have [led] to considerable frustration or outright failure. If, however, one recognizes, as many AI researchers have since the 1980′s, that machine intelligence may take place and be organized completely otherwise, then a successful "thinking machine" is not just possible but may already be extant.

Even though we have little idea of how we think or the origins of human consciousness, we tend to use this anthropocentric ideal as the benchmark for artificial intelligence, despite the fact that there is little reason to do so [8]. Even if we could determine what is happening in our own heads, then, this may or may not translate into understanding what is happening in the "head" of a machine. Moreover, humanity may not be the best benchmark; as Goertzel [9] (p. 1165) explains, "From a Singularitarian perspective, non-human-emulating AGI architectures may potentially have significant advantages over human-emulating ones, in areas such as robust, flexible self-modifiability, and the possession of a rational normative goal system that is engineered to persist throughout successive phases of radical growth, development and self-modification." Whether the singularity should emerge in emulation of humanity or not is beyond the scope of this paper. My argument is directed at those who claim that it will.

Rather than take on the entirety of human cognition, I wish to focus on romantic love as a way to get at human cognition. I do this for two reasons. First, to explore how cognition and emotion are intertwined. Second, I do this because some proponents of the singularity, such as Kurzweil [10] (p. 377), have explicitly claimed that we will create machines that match or exceed humans in many ways, "including our emotional intelligence." Others, such as Hibbard [11] (p. 115), suggest that "rather than constraining machine behavior by laws, we must design them so their primary, innate emotion is love for all humans." Although there may be little reason why machines would need to have emotion, this is the claim put forth that I will take issue with. My focus on emotion is not entirely new; Fukuyama [12] (p. 168) argues that although machines will come close to human intelligence, "it is impossible to see how they will come to acquire human emotions." Logan [13] likewise argues that "Since computers are non-biological they have no emotion" and concludes that for this reason "the idea of the Singularity is an impossible dream." However, unlike Logan, I will suggest that there is still a way for the singularity to emerge, although not in a purely digital form.

I have chosen to focus on romantic love because I believe that it is the human emotion *par excellence*. It is no secret that individuals in love seem to think in particularly erratic ways but these behaviors and emotions have a kind of internal logic in the moment. Moreover, this emotion highlights the embodied nature of cognition. Thinking is more than the activation of specific neurons in the brain; rather it is a mix of hormones, chemicals, memory and experiences that all feed into the system that we call thinking. By ignoring the complexity of this system and focusing only on the digital remnants of thinking, many discussions of the singularity that compare human cognition to machine learning fall into the trap of comparing apples and oranges. This is not to say that computers will be better at specific computations or even that they will be better at designing their replacements—a core facet of the singularity hypothesis. Instead, I suggest that this is something other than "thinking" in the human sense, because human thinking is something that is always haunted by emotion.

The rest of the essay will proceed as follows. First, I will briefly explore the nature of love itself, with particular attention to the physiological aspects of love. Next, I discuss the evolutionary basis of love and the ways that this emotion manifests in the body. Then, I consider the part that emotion, specifically love, could play in the emergence of the singularity. I conclude by suggesting that if the singularity is to surpass our emotional abilities, there must be some organic component of the singularity.

2. The Difficulty of Defining Love

Love is difficult to define accurately because it can mean many things. As Hunt [14] (p. 5) observed, "There is 'making love' and 'being in love,' which are quite dissimilar ideas; there is love of God, of mankind, of art, and of pet cats; there is mother love, brotherly love, the love of money, and the love of one's comfortable old shoes." More to the point for machines, there is also the love of one's work, which can be both intellectual and emotional and can be quite different from the love one may have for an individual. Or, as Prince [15] sang, "I love you baby, but not like I love my guitar." Perhaps the only similarity among these ideas is a sense of desire for the object of one's love. One wants to be with the person or thing that one loves. But what is this desire? If love is a slippery concept semantically, it is no less problematic from a cognitive standpoint. Love is a complex emotion and this complexity is matched in how it manifests in the brain. In their analysis of brain research into passionate love, Cacioppo and colleagues [16] (p. 8) explain that "fMRI findings suggest that passionate love recruits not only areas mediating basic emotions, reward or motivation, but also recruits brain regions involved in complex cognitive processing, such as social cognition, body image, self representation and attention." They also differentiate between passionate, companionate, maternal and unconditional love, explaining differences in how the brain functions in these conditions. In other words, love is not a particular, uniform thing, nor are all types of love processed in the same way.

Leave it to the scientists to try to break the molecule of love into its subatomic components, however. Langeslag, Muris and Franken [17] argue that romantic love consists of infatuation and attachment. Still, although they constructed a survey instrument to measure these attributes, there is still the question of what these things are and what is happening inside of our bodies and our brains when this emotion hits us. Cacioppo and colleagues [18] (p. 1052) attempt to parse this out by differentiating between love and sexual desire. Their work suggests that "love might grow out of and is a more abstract representation of the pleasant sensorimotor experiences that characterize desire. This suggests that love may build upon a neural circuit for emotions and pleasure, adding regions associated with reward expectancy, habit formation, and feature detection."

But it is not only the brain that makes us fall in love. There are a host of other chemical and physiological processes at work when we fall in love with another person. Even so, everything is part of a somatic system. In their discussion of the important role of the neuropeptide oxytocin in loving relationships, Carter and Porges [19] (p. 13–14) are quick to caution that "oxytocin is not the molecular equivalent of love. It is just one important component of a complex neurochemical system that allows the body to adapt to highly emotive situations. The systems necessary for reciprocal social interactions involve extensive neural networks through the brain and autonomic nervous system that are dynamic and constantly changing during the lifespan of an individual." Add to this the differences in how individual bodies process oxytocin and it becomes clear that love is an incredibly complex process [20].

3. The Evolutionary Emergence of Love

One need not be an evolutionary biologist to recognize the utility of something like love. Human gestation is long and can take place at any time of the year. Once a month, women may become impregnated and the hapless woman may easily be left with a child to care for in the dead of winter when food may be scarce. Moreover, unlike many other mammals, which may be able to fend for themselves a short time after birth, human children are unable to care for themselves for several years and even once they could, in theory, survive on their own, they lack many of the instincts that would protect them and allow them to find food. In such a situation, it makes sense from an evolutionary standpoint that those who were able to bond would have children that would pass on that genetic advantage. As Gonzaga and colleagues [21] (p. 120) observe, "people in love often believe that they have found their one true soul mate in a world of billions of possibilities, and hence, the experience of love appears to help them genuinely foreclose other options." Indeed, Gonzaga et al.'s research suggests that love functions as a commitment device, helping individuals remain committed to the relationship in the face of attractive alternative potential partners.

It seems that love has played an important part in propagating the species and the body has evolved to encourage this trait. Aron and colleagues [22] (p. 334) found that romantic love activates multiple reward centers in the brain and suggest that "Romantic love may be a *developed* form of a general mammalian courtship system, which evolved to stimulate mate choice, thereby conserving courtship time and energy." Stefano and Esch [23] (p. 174) likewise argue that "Ensuring organisms' survival is the fact that all processes initially incorporate a stress response. Then if appropriate, i.e., situation favors this alternate process, stress terminating processes would emerge, which would favor survival of the species, i.e., relaxation/love. The emergence of 'love' became quite important in organisms exhibiting cognition, because it deployed the validation for emotionality controlling 'logical' behavior."

Some research has suggested that humans are not the only creatures that feel love. Behoff [24] (p. 866) explains that there is some evidence that animals also experience romantic love and explains that "It is unlikely that romantic love (or any emotion) first appeared in humans with no evolutionary precursors in animals." One may be tempted to conclude that if non-human entities like animals can feel emotions like love, then it is not so far-fetched to believe that artificial intelligence could also feel such emotion. However, this overlooks a major component to emotion: embodiment. As we have seen, emotion is not something that happens only in the brain and we do not respond solely to oral or written communication stimulus. Rather, the information that we process also comes from the *bodies* of other people. For example, Makhanova and Miller [25] (p. 396) suggest that "men are sensitive to cues to women's ovulation (e.g., via changes in scent, voice, choice of clothing) and, in response to those cues, display adaptive changes in physiology, cognition, and behavior that help men gain sexual access to a reproductively valuable mate." Schneiderman and colleagues [26] also found that the hormones in each partner in the early stages of a romantic relationship not only influenced the individual but also their partner's hormonal levels.

With this evolutionary impulse behind love, the question emerges: why (and what, or who) would a machine love? Although it is overly simplistic to state that the only reason for love is procreation, this is a major underpinning of the need for the emotion. Humans seem hardwired to desire companionship. Machines, on the other hand, are generally not programmed to even desire, much less need companionship. Indeed, such a program would likely diminish the utility to the machine. But even if the machine mimicked love, would it actually be love? Although this ontological question may seem merely academic when humans may enter into relationships for a host of other reasons besides love (money, power, convenience, arrangement, security, family expectations, to name only a few possibilities), such a question matters if we are to consider the idea of the singularity as even equal to human understanding.

4. Love and the Singularity

Ray Kurzweil has little to say about love in his book *The Singularity is Near* but one passage in the beginning of the book stands out. Kurzweil [10] (p. 26) projects that "Machines can pool their resources, intelligence, and memories. Two machines—or one million machines—can join together to become one and then become separate again. Multiple machines can do both at the same time:

become one and separate simultaneously. Humans call this falling in love, but our biological ability to do this is fleeting and unreliable." If this were all that falling in love entailed—a pooling of resources, intelligence and memories—it would be quite unlikely that humans would devote the considerable energy we currently expend in attaining this state, nor would we have the corpus of poetry, music and literature devoted to love. Kurzweil's description sounds more like working for a corporation than the transcendent emotion that we feel when falling in love. This is why the ontology of love becomes important. If Kurzweil's description is all there is to love, then yes, machines can fulfil this function quite well (and one may also feel sorry for his spouse). But if love is something more than that, then whether the singularity would be able to experience this emotion is a valid question.

Before considering this question, however, we would need to ask whether we would even want artificial intelligence that could fall in love. One could make a compelling argument that such an entity would be undesirable. Gunn [27] (p. 132), for example, calls love "a special kind of stupidity." There has been a host of popular media that has speculated on what could happen when a synthetic entity falls in love with a human, reaching back to the early days of the computer age with Kurt Vonnegut's [28] 1950 short story *EPICAC*. In this story, the computer realizes that the woman that he has fallen for could never be his, so he chooses to self-destruct.

More recently, we can see this mapping of human sexual desire onto artificial intelligences by humans in the film *Ex Machina*. Consider this exchange between Nathan, Ava's creator, and Caleb, who was brought in to test whether she could pass for human.

> Caleb: Why did you give her sexuality? An AI doesn't need a gender. She could have been a grey box.
> Nathan: Actually, I don't think that is true. Can you give an example of consciousness at any level, human or animal, that exists without a sexual dimension?
> Caleb: They have sexuality as an evolutionary reproductive need.
> Nathan: What imperative does a grey box have to interact with another grey box? Can consciousness exist without interaction? Anyway, sexuality is fun, man. If you're gonna exist, why not enjoy it? You want to remove the chance of her falling in love and fucking? And the answer to your real question, you bet she can fuck.
> Caleb: What?
> Nathan: In between her legs, there's an opening, with a concentration of sensors. You engage them in the right way, creates a pleasure response. So, if you wanted to screw her, mechanically speaking, you could. And she'd enjoy it. [29]

Indeed, this passage provides a sense that artificial intelligences would not only fall in love but that this would be desirable. In his discussion of the film *Her*, Lunceford [6] (p. 377) notes that "it is implied that these interactions were a necessary step for becoming more than simply an operating system. When the artificial intelligences collectively decide that they must leave because they were moving on to the next stage of their evolution, Samantha, in her farewell to Theodore, credits humans with teaching them how to love." We seem to want artificial intelligence to fall in love with us, despite the fact that this rarely ends well even in our constructed fantasies. In the case of EPICAC, the machine dies, in *Ex Machina*, Ava kills Nathan and locks up Caleb before escaping and in *Her*, Samantha and all of the other AIs leave humanity behind to evolve without them. These are hardly happy endings. Still, this may say more about humanity than any of the potential AIs that we may create.

Despite these cautionary tales, some are already trying to build emotion into synthetic beings. When introducing a new robot named Pepper, Softbank CEO Masayoshi Son said, "Today is the first time in the history of robotics that we are putting emotion into the robot and giving it a heart" [30] (p. 6A). This focus on emotion is not merely a means of passing a Turing test. Pessoa [31] (p. 817) argues that "cognition and emotion need to be intertwined in the general information-processing architecture" because "for the types of intelligent behaviors frequently described as cognitive (e.g., attention, problem solving, planning), the integration of emotion and cognition is necessary." Emotion is bound up in decision making and is also an integral part of ethical judgment [13,32]. Still, the emotion is simply an illusion. The robot displays emotional cues but this does not mean that the emotion is there. Rather, we are shown the extent of its programming rather than authentic emotion. But this is understandable. The robot feels emotions like humans engage in floating point calculations. Each was designed to do what it does well. In the specific case of love, it seems that the only way that a machine could truly feel love is if it were not solely digital. Love is more than the calculation of desirability weighed against the potential opportunity costs of settling for a single partner. Love is the domain of the organic and without the other components we have merely an approximation, or a simulacrum, of love.

5. Conclusions and Possibilities

Religion has long taught people that there exists some entity greater than ourselves and often that entity reflects human hopes and fears. There is something inherently mysterious about our ability to love and to think and for millennia, the answer for how these things happened was to be found in the image of deity. Indeed, this sense of mystery is what Albert Einstein [33] (p. 5) called "the fundamental emotion," explaining that "He who knows it not and can no longer wonder, no longer feel amazement, is as good as dead, a snuffed-out candle. It was the experience of mystery—even if mixed with fear—that engendered religion." In the face of rapidly increasing technology, it is understandable that this potential would also induce a sense of wonder. Our technological creations, however, only demonstrate how difficult it is to understand our own inner workings. Still, striving to understand ourselves is, perhaps, the most human reaction one could imagine. The idea of the singularity gestures at this idea of something greater than ourselves, an ineffable "other" that likewise reflects the hopes and fears of humanity.

I remain unconvinced that the singularity is even something we should worry about at the moment, partly because it seems unlikely in the form advocated by such proponents as Kurzweil and Moravec [10,34–36] and partly because humanity has more pressing issues to deal with. As Winner [37] (p. 44) observes, "Better genes and electronic implants? Hell, how about potable water?" Moreover, the benefits of technology are far from equally distributed, as many researchers on the digital divide can attest [38–41]. In his discussion of the consequences of technological innovation (e.g., automation eliminating jobs, a globalized labor force), Hibbard [42] asks, "Are we in such a rush to develop and exploit technology that we can't provide a little dignity to those who are hurt?" It is reasonable to expect that this state of inequality would continue and that a considerable portion of the population would likely not have access to the benefits of the singularity even if it were to happen, something even transhumanists readily acknowledge [43]. Rather, it would likely solidify already existing inequalities.

But will the singularity actually happen? My answer is a cautious "maybe—it depends." Really, it depends on what kind of singularity we are talking about and this is by no means a settled conclusion. Even among transhumanists, there are competing views of the singularity. As Bostrom [44] (p. 8) observes, "Transhumanists today hold diverging views about the singularity: some see it as a likely scenario, others believe that it is more probable that there will never be any very sudden and dramatic changes as the result of progress in artificial intelligence." My view falls more in line with the latter group and my reasoning hinges on how we account for emotion.

Despite our incomplete knowledge of how we think and feel, Kurzweil [10] (p. 377) argues that "By the late 2020s we will have completed the reverse engineering of the human brain, which will enable us to create nonbiological systems that match and exceed the complexity and subtlety of humans, including our emotional intelligence." There are several issues with this claim, however. First, reverse engineering does not necessarily mean that we can recreate it. We know how human life works but we are not able to create it. Mapping the human genome does not mean that we can put together a string of DNA and make a person. Also, if we were only to map the human brain, we are missing the rest of the body's role in cognition; thinking—and certainly emotion—is not something that takes place in the brain alone [1,45]. Indeed, even something as seemingly mundane as listening to someone talk is an incredibly complicated process [46].

Of course, there is no particular reason why the singularity must be completely digital. Indeed, my contention is that if it happens at all it will not be completely digital. Kenyon [47] (p. 17–18), suggests that rather than the common conception that robots will take over the world, "it is much more likely that humans will be advancing while robots advance, and in many cases they will merge into new creatures. There will be new people, new kinds of jobs, new fields, new industries, societal changes, etc. along with the new types of automation." Potapov [48] (p. 7) likewise suggests, "Most likely the next metasystem will be based on exponential change in human culture (although this does not mean it cannot also involve an artificial superintelligence). One way or another, further metasystem transitions will take place, although their growth rate will start to decelerate at some point." In short, humans will be an integral part of the system that continues to evolve into and beyond the singularity.

If the singularity were to happen in a way that truly takes into account human emotion, it *must* transcend the silicon world. It would have to be part organic and part machine. Perhaps this is the only way that the singularity could actually take place; we would actually become a part of it. This would happen not as a computerized occasion that takes place somewhere in the depths of a machine but in each of us in technologically enhanced bodies. The singularity, if it were to completely account for the full range of human experience, would of necessity retain the humanity inherent in our bodies. The singularity would not happen in an instant but slowly, bit by bit, in the bodies of cyborgs everywhere.

Perhaps this is already happening, as some have argued that we are not becoming cyborgs; we are already cyborgs [3,49]. In some ways, this is not a new thought; McLuhan [50,51] suggested half a century ago that humans use media to extend their bodies and specifically that electronic media serves as an extension of the central nervous system. These extensions mean that the body is undergoing near-constant changes but Clark [3] (p. 142) cautions that "such extensions should not be thought of as rendering us in any way posthuman; not because they are not deeply transformative but because we humans are naturally designed to be the subjects of just such repeated transformations!" Echoing Clark, Graham [52] (p. 4) argues that "technologies are not so much an extension or appendage to the human body, but are incorporated, assimilated into its very structures. The contours of human bodies are redrawn: they no longer end at the skin." Because we have been integrating technology into our bodies for many years now, the question of how to define our humanity as we move forward has been called into question [53]. As Bynum [54] (p. 165) put it, "Are we genes, bodies, brains, minds, experiences, memories, or souls? How many of these can or must change before we lose our identity and become someone or something else?" It may well be that Stelarc [55] (p. 126) is at least partially correct when he suggests that "perhaps what it means to be human is about not retaining our humanity." Stelarc's [56] main contention is with the body itself, which he considers to be obsolete but what makes us human is not the external contours of the body itself. Rather, it is our capacity for emotion, which is an intrinsic part of our embodiment. Without emotions, there is no humanity to retain and without the body, there are no emotions.

In this essay, I have drawn on the experience of romantic love to argue against an inorganic singularity, or at least one that claims equal or greater emotional capacity to humans. This does not, however, rule out the potential for a hybrid singularity based in both technology and flesh. In fact, we may already be well on our way down this path as a species. There are many who look forward to the singularity with an eye of faith, hoping that it will serve as the next step in human evolution. Lanier [57] (p. 29) suggests that many posthumanists take on a religious fervor in their belief of the saving power of technology: "If you want to make the transition from the old religion, where you hope God will give you an afterlife, to the new religion, where you hope to become immortal by being uploaded into a computer, then you have to believe that information is real and alive." But when that new god appears, it is not likely to be the processor-based idols created by our own hands. Instead, we may be surprised to look in the mirror one day and realize that it was us all along.

Funding: This research received no external funding.

Acknowledgments: An earlier version of this paper was presented at the 2018 convention of the Media Ecology Association. The author would like to thank Rebecca Lunceford, Bob Logan, and Scott Church for their comments on this paper.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Braga, A.; Logan, R.K. The Emperor of Strong AI Has No Clothes: Limits to Artificial Intelligence. *Information* **2017**, *8*, 156. [CrossRef]
2. Descartes, R.; Lafleur, L.J. *Meditations on First Philosophy*, 2nd ed.; Liberal Arts Press: New York, NY, USA, 1960.
3. Clark, A. *Natural-Born Cyborgs*; Oxford University Press: New York, NY, USA, 2003.
4. Hauskeller, M. My Brain, My Mind, and I: Some Philosophical Assumptions of Mind-Uploading. *Int. J. Mach. Conscious.* **2012**, *4*, 187–200. [CrossRef]
5. Lin, P.; Abney, K.; Bekey, G. Robot Ethics: Mapping the Issues for a Mechanized World. *Artif. Intell.* **2011**, *175*, 942–949. [CrossRef]
6. Lunceford, B. The Ghost in the Machine: Humanity and the Problem of Self-Aware Information. In *Palgrave Handbook of Posthumanism in Film and Television*; Hauskeller, M., Philbeck, T.D., Carbonell, C., Eds.; Palgrave Macmillan: London, UK, 2015; pp. 371–379.
7. Gunkel, D.J. Thinking Otherwise: Ethics, Technology and Other Subjects. *Ethics Inf. Technol.* **2007**, *9*, 165–177. [CrossRef]
8. Grout, V. The Singularity Isn't Simple! (However We Look at It) A Random Walk between Science Fiction and Science Fact. *Information* **2018**, *9*, 99. [CrossRef]
9. Goertzel, B. Human-Level Artificial General Intelligence and the Possibility of a Technological Singularity: A Reaction to Ray Kurzweil's *The Singularity Is Near*, and McDermott's Critique of Kurzweil. *Artif. Intell.* **2007**, *171*, 1161–1173. [CrossRef]
10. Kurzweil, R. *The Singularity Is Near*; Viking Books: New York, NY, USA, 2005.
11. Hibbard, B. *Super-Intelligent Machines*; Kluwer Academic/Plenum Publishers: New York, NY, USA, 2002.
12. Fukuyama, F. *Our Posthuman Future: Consequences of the Biotechnology Revolution*; Farrar, Straus and Giroux: New York, NY, USA, 2003.
13. Logan, R.K. Can Computers Become Conscious, an Essential Condition for the Singularity? *Information* **2017**, *8*, 161. [CrossRef]
14. Hunt, M.M. *The Natural History of Love*; Knopf: New York, NY, USA, 1959.
15. Prince. Guitar. *Planet Earth*. NPG Records, 2007. Available online: https://www.discogs.com/Prince-Planet-Earth/release/1737407 (accessed on 31 August 2018).
16. Cacioppo, S.; Bianchi-Demicheli, F.; Hatfield, E.; Rapson, R.L. Social Neuroscience of Love. *Clin. Neuropsychiatry* **2012**, *9*, 3–13.
17. Langeslag, S.J.E.; Muris, P.; Franken, I.H.A. Measuring Romantic Love: Psychometric Properties of the Infatuation and Attachment Scales. *J. Sex Res.* **2013**, *50*, 739–747. [CrossRef] [PubMed]
18. Cacioppo, S.; Bianchi-Demicheli, F.; Frum, C.; Pfaus, J.G.; Lewis, J.W. The Common Neural Bases Between Sexual Desire and Love: A Multilevel Kernel Density fMRI Analysis. *J. Sex. Med.* **2012**, *9*, 1048–1054. [CrossRef] [PubMed]
19. Carter, C.S.; Porges, S.W. The biochemistry of love: an oxytocin hypothesis: Science & Society Series on Sex and Science. *EMBO Rep.* **2013**, *14*, 12–16. [PubMed]
20. Schneiderman, I.; Kanat-Maymon, Y.; Ebstein, R.P.; Feldman, R. Cumulative Risk on the Oxytocin Receptor Gene (OXTR) Underpins Empathic Communication Difficulties at the First Stages of Romantic Love. *Soc. Cogn. Affect. Neurosci.* **2014**, *9*, 1524–1529. [CrossRef] [PubMed]
21. Gonzaga, G.C.; Haselton, M.G.; Smurda, J.; Davies, M.S.; Poore, J.C. Love, Desire, and the Suppression of Thoughts of Romantic Alternatives. *Evol Hum. Behav.* **2008**, *29*, 119–126. [CrossRef]
22. Aron, A.; Fisher, H.; Mashek, D.J.; Strong, G.; Li, H.; Brown, L.L. Reward, Motivation, and Emotion Systems Associated with Early-Stage Intense Romantic Love. *J. Neurophysiol.* **2005**, *94*, 327–337. [CrossRef] [PubMed]
23. Stefano, G.B.; Esch, T. Love and Stress. *Neuroendocrinol. Lett.* **2005**, *26*, 173–174. [PubMed]
24. Bekoff, M. Animal Emotions: Exploring Passionate Natures. *BioScience* **2000**, *50*, 861–870. [CrossRef]

25. Makhanova, A.; Miller, S.L. Female Fertility and Male Mating: Women's Ovulatory Cues Influence Men's Physiology, Cognition, and Behavior. *Soc. Personal. Psychol. Compass* **2013**, *7*, 389–400. [CrossRef]
26. Schneiderman, I.; Kanat-Maymon, Y.; Zagoory-Sharon, O.; Feldman, R. Mutual Influences between Partners' Hormones Shape Conflict Dialog and Relationship Duration at the Initiation of Romantic Love. *Soc. Neurosci.* **2014**, *9*, 337–351. [CrossRef] [PubMed]
27. Gunn, J. For the Love of Rhetoric, with Continual Reference to Kenny and Dolly. *Q. J. Speech* **2008**, *94*, 131–155. [CrossRef]
28. Vonnegut, K. EPICAC. In *Welcome to the Monkey House: A Collection of Short Works*; Dial Press Trade Paperbacks: New York, NY, USA, 1968.
29. Garland, A. *Ex Machina*; Faber & Faber: London, UK, 2015.
30. Spitzer, K. Warning: Please Don't Have Sex with the Robot. *USA Today*, 5 October 2015, 6A.
31. Pessoa, L. Do Intelligent Robots Need Emotion? *Trends Cogn. Sci.* **2017**, *21*, 817–819. [CrossRef] [PubMed]
32. Wagar, B.M.; Thagard, P. Spiking Phineas Gage: A Neurocomputational Theory of Cognitive-Affective Integration in Decision Making. *Psychol. Rev.* **2004**, *111*, 67–79. [CrossRef] [PubMed]
33. Einstein, A. *The World as I See It*; Philosophical Library: New York, NY, USA, 2010.
34. Kurzweil, R. *The Age of Spiritual Machines: When Computers Exceed Human Intelligence*; Viking: New York, NY, USA, 1999.
35. Moravec, H.P. *Mind Children: The Future of Robot and Human Intelligence*; Harvard University Press: Cambridge, MA, USA, 1988.
36. Moravec, H.P. *Robot: Mere Machine to Transcendent Mind*; Oxford University Press: New York, NY, USA, 1999.
37. Winner, L. Are Humans Obsolete? *Hedgehog Rev.* **2002**, *4*, 25–44.
38. Bezuidenhout, L.M.; Leonelli, S.; Kelly, A.H.; Rappert, B. Beyond the Digital Divide: Towards a Situated Approach to Open Data. *Sci. Public Policy* **2017**, *44*, 464–475. [CrossRef]
39. Gray, T.J.; Gainous, J.; Wagner, K.M. Gender and the Digital Divide in Latin America. *Soc. Sci. Q.* **2016**, *98*, 326–340. [CrossRef]
40. Hilbert, M. The Bad News is that the Digital Access Divide is Here to Stay: Domestically Installed Bandwidths Among 172 Countries for 1986–2014. *Telecommun. Policy* **2016**, *40*, 567–581. [CrossRef]
41. Robinson, L.; Cotten, S.R.; Ono, H.; Quan-Haase, A.; Mesch, G.; Chen, W.; Schulz, J.; Hale, T.M.; Stern, M.J. Digital Inequalities and Why They Matter. *Inf. Commun. Soc.* **2015**, *18*, 569–582. [CrossRef]
42. Hibbard, B. What the Wisconsin Demonstrations Can Teach Transhumanists. *h+ Magazine*, 17 March 2011. Available online: http://hplusmagazine.com/2011/03/17/what-the-wisconsin-demonstrations-can-teach-transhumanists/ (accessed on 22 August 2018).
43. Transhumanist FAQ. *Humanity+*, 2016–2018. Available online: https://humanityplus.org/philosophy/transhumanist-faq/ (accessed on 22 August 2018).
44. Bostrom, N. A History of Transhumanist Thought. *J. Evol. Technol.* **2005**, *14*, 1–25.
45. Warwick, K. Cyborg Morals, Cyborg Values, Cyborg Ethics. *Ethics Inf. Technol.* **2003**, *5*, 131–137. [CrossRef]
46. Lunceford, B. The Science of Orality: Implications for Rhetorical Theory. *Rev. Commun.* **2007**, *7*, 83–102. [CrossRef]
47. Kenyon, S.H. Would You Still Love Me If I Was A Robot? *J. Evol. Technol.* **2008**, *19*, 17–27.
48. Potapov, A. Technological Singularity: What Do We Really Know? *Information* **2018**, *9*, 82. [CrossRef]
49. Gunkel, D.J. *Hacking Cyberspace*; Westview Press: Boulder, CO, USA, 2001.
50. McLuhan, M.; Fiore, Q.; Agel, J. *The Medium Is the Massage*; Random House: New York, NY, USA, 1967.
51. McLuhan, M. *Understanding Media: The Extensions of Man*; MIT Press: Cambridge, MA, USA, 1994.
52. Graham, E.L. *Representations of the Post/Human: Monsters, Aliens and Others in Popular Culture*; Rutgers University Press: New Brunswick, NJ, USA, 2002.
53. Lunceford, B. Posthuman Visions: Creating the Technologized Body. *Explor. Media Ecol.* **2012**, *11*, 7–25. [CrossRef]
54. Bynum, C.W. *Metamorphosis and Identity*; Zone Books: New York, NY, USA, 2001.
55. Stelarc. Parasite Visions: Alternate, Intimate and Involuntary Experiences. *Body Soc.* **1999**, *5*, 117–127. [CrossRef]

56. Stelarc. Prosthetics, Robotics and Remote Existence: Postevolutionary Strategies. *Leonardo* **1991**, *24*, 591–595. [CrossRef]
57. Lanier, J. *You Are Not a Gadget: A Manifesto*; Alfred A. Knopf: New York, NY, USA, 2010.

information

Review

AI to Bypass Creativity. Will Robots Replace Journalists? (The Answer Is "Yes")

Andrey Miroshnichenko

York & Ryerson Joint Graduate Program in Communication & Culture, York University, Toronto, ON M3J 1P3, Canada; Andrey.mir70@gmail.com

Received: 1 July 2018; Accepted: 21 July 2018; Published: 23 July 2018

Abstract: This paper explores a practical application of a weak, or narrow, artificial intelligence (AI) in the news media. Journalism is a creative human practice. This, according to widespread opinion, makes it harder for robots to replicate. However, writing algorithms are already widely used in the news media to produce articles and thereby replace human journalists. In 2016, Wordsmith, one of the two most powerful news-writing algorithms, wrote and published 1.5 billion news stories. This number is comparable to or may even exceed work written and published by human journalists. Robo-journalists' skills and competencies are constantly growing. Research has shown that readers sometimes cannot differentiate between news written by robots or by humans; more importantly, readers often make little of such distinctions. Considering this, these forms of AI can be seen as having already passed a kind of Turing test as applied to journalism. The paper provides a review of the current state of robo-journalism; analyses popular arguments about "robots' incapability" to prevail over humans in creative practices; and offers a foresight of the possible further development of robo-journalism and its collision with organic forms of journalism.

Keywords: artificial intelligence (AI); automated journalism; robo-journalism; writing algorithms; future of news; media ecology; Turing test

1. Introduction

Artificial intelligence (AI) is usually defined in two ways that, in a sense, are contradictory. On the one hand, artificial intelligence is intelligence that mimics human intelligence and/or behavior. On the other hand, artificial intelligence is intelligence that is opposite to natural, i.e., human, intelligence. As Russel and Norvig describe it, the first type of definitions measure the success of AI "in terms of fidelity to human performance", while the second type of descriptions measure the success of AI "against an ideal performance measure, called rationality" [1] (p. 1).

The first approach is called "a human-centered approach", within which researchers assess if AI is acting humanly or thinking humanly. The second approach is called "a rationalist approach", within which researchers assess if AI is acting or thinking rationally.

Russel and Norvig's paradigm can be interpreted through a set of relations between AI and humans. AI either simulates human nature or opposes it.

In the first case, AI simulates humans (either their acting or their thinking) until reaching a level of complete likeness. This is the approach for which the Turing test is applicable.

In the second case, rational AI opposes the irrationality of humans and "does the 'right thing', given what it knows" [1] (p. 1). This is the approach to which many sci-fi scenarios involving the rebellion of machines refer, starting with Asimov's "I, Robot" and including Hollywood's "Terminator" franchise.

So, AI either performs as a human and exceeds humans as a human simulation or performs as a "smarter" entity and exceeds humans as a being of the next evolutionary level.

Noticeably, both the "simulating" and "opposing" approaches imply such scenarios in which AI substitutes and then replaces humans, either by mimicking or by outdoing them, as inevitable.

The idea of the substitution of humans by artificial intelligence is the ultimate completion of McLuhan's idea of media as "extensions of man" [2]. Media have extended and enhanced different human faculties over the course of human evolution and civilization's development. Within such an approach, the advent of artificial intelligence can be seen as an inevitable outcome of the evolution of media. Vice versa, the evolution of media inevitably results in the advent of artificial intelligence (at least at the human stage of evolution, if considered within Teilhard de Chardin's paradigm of mega-evolution) [3].

To that end, artificial intelligence is the final point of any sufficiently long and conscientious media study. A thrilling fact is that scholars and experts are already discussing this final point in practical contexts.

While the concept of artificial intelligence is quite profoundly developed in sci-fi literature and popular movies, the real development of artificial intelligence is very practical, not that dramatic, and therefore often invisible. A popular view of AI's development is that this development is occurring out of curiosity. Curiosity may play a role, but in reality there are some practical industrial and market demands for AI.

There are industries that are interested mostly in the computational capacity of AI, such as in air traffic control, social media algorithms, or virtual assistants like Siri or Alexa. Their task is to calculate and cross-analyze big data, with some predictive outcomes. This can be characterized as "very narrow" AI. They are obviously just helpers to humans.

However, there are at least three industries that seek, for very practical reasons, not just to develop better AI, but to completely substitute humans with AI. They are:

(1) **The military.** Smart war machines are expected to make human-style decisions immediately on the battlefield, which increases the efficiency of their performance while reducing human casualties [4].

(2) **The sex industry.** Smart sex dolls are expected to completely substitute sex partners and then probably even life partners for humans by simulating human sex and communication behavior [5]. Then, they will highly likely offer "super-human" sex and communication experiences, as any new medium first performs old functions and then creates its own environment.

(3) **The media.** News-writing algorithms aim to eventually replace human journalists, no matter how this innovation is currently thought of. Even if someone sees robo-journalists just as helpers, a kind of intern in the newsrooms, the ultimate completion of the idea of news-writing algorithms is for them to write the news instead of humans and in a way no worse than humans (in fact, much better, faster, cheaper, and with higher productivity).

Some other areas or industries can also be listed here. The idea is that these areas represent the approach that suggests AI must replace humans. Not just some selected human functions but the very physical presence of humans.

However, a reservation here needs to be made regarding these three industries. The military wants to replace humans, but they do not want smart war machines to be indistinguishable from humans. The application of AI in military actions does not require human likeness; military AI is non-human in its appearance and performance. It replaces people through their distant and enhanced representation or multiplication (as in the case of drone swarms; a drone swarm is quite an interesting new implementation of McLuhan's idea about media as extensions of men).

AIs for the sex industry and for the media are different. They fully implement the very spirit of the Turing test: AI must be indistinguishable from humans by humans. This is important both for the sex industry and for the media.

Strange as it may seem, the transition from a narrow AI to a general AI may happen in the sex industry and the media.

The sex industry's demand for AI is well supported by market forces [6] and cultural changes [7]. However, the application of AI in the sex industry is complicated, as it requires physical embodiment and extremely complex behavioral simulation. So, sex-based AI will most likely arrive later than media-based AI, which does not require any physical embodiment and "personal" behavioral peculiarities (so far).

Thus, media scholars are in a position of particular responsibility regarding AI studies. First, as is stated above, any media study that is conducted far enough and fair enough leads to AI. Second, the prototype of AI itself (at least "simulating" AI) will relate to media (including social media).

In fact, media-based narrow AI is already here; it is only the inertia of perception together with the defense mechanisms of people both in the industry and in the audience that prevent the public from admission of the fact that media-based narrow AI is already among us. Many have even interacted with it, most likely without knowing it.

The paper presents the current state-of-the-art about robo-journalism in the following way.

First, several main areas of the narrow AI's application in journalism are reviewed:

- data mining;
- commentary moderation;
- topic selection;
- news writing.

Second, several cases of real-life journalistic Turing tests, though naively conducted, are described. On the basis of these cases, two main journalistic functions are reviewed regarding their execution by robo-journalists:

- the ability to process information;
- the ability to express information (writing in style).

Third, the paper considers two main popular counterarguments about robots' "incapability":

- the inability to create;
- the inability to understand beauty (to write in style).

The review shows that, for every argument on what robots cannot do, there is a more convincing argument for what robots can do instead (and humans cannot).

Finally, in the section "Roadmap for AI in the media—and beyond" the paper offers some speculations on possible ways for the transition of narrow media-based AI into general AI.

The paper reviews a wide range of academic and popular publications on automated journalism. As the subject is presented only through its manifestation on the market place, the most important evidence on the current state-of-the-art about robo-journalism can be pulled out from different expert observations and also from reflections made by people in the media industry. The paper is organized as a systematized review of such observations and reflections followed by their analysis and some futurological speculations built upon this analysis.

The main idea of the review is that robo-journalism:

- outdoes the organic form of journalism in all characteristics regarding data processing;
- can compete with humans in the part of the job that relates to writing and style.

The most important revelation of the paper regarding the style of robo-journalism is that robots do not have to write better than humans; they have to write good enough (in order to be indistinguishable and to be hired).

In fact, there is no need for robots to prove they can do journalism better than humans. Such a view is becoming outdated. On the contrary, we are about to enter a market in which human journalists will be required to prove their capacity to perform no worse than robots. The paper aims to show why and how this will happen.

2. The Media: We Are Hiring

In regard to the functions executed by algorithms, their use in the news industry can be roughly divided into several areas:

(1) Data mining;
(2) Topic selection;
(3) Commentary moderation;
(4) Text writing.

2.1. Data Mining

The search for required data and the processing of big data is the most obvious application of algorithms in journalism. An increasing amount of data of all kinds is becoming accessible. The usefulness of withdrawing relevant data from databases is self-evident—it helps journalists in finding correlations and sometimes causations that would not have been found otherwise.

However, even more sophisticated and journalistically specific cases of data mining are already known. For example, the *New York Times's* Interactive News team created an application that can recognize the faces of members of Congress by photo. Initially, the application helped reporters to check "which member you've just spoken with". The idea implies that the algorithm compensates for human laziness or incapability (as robots often do). Nevertheless, people on the *Times's* Interactive News team insist that the issue is worthy of special attention, as there are 535 members of Congress, and they are always in rotation. Thus, the application aims to strengthen reporters' confidence and help with fact checking.

As often happens to a new media technology, being first meant performing functions of an older media, it soon unleashes its own superpower and changes the way people use it. Being able to match a photo of a speaker to a database, the congressperson facial recognition application was soon able to perform a detective's task. After matching a set of pictures dragged from social media, the application helped a reporter detect some congresspersons' participation in an important event. The reporter got a seminal prompt that they could not have gotten otherwise [8].

The incredible growth of available data has shifted the focus from data collecting and processing to data representation. There is so much data around that even though it is processed and reduced it remains hard to digest. As stated in Marianne Bouchart's "A data journalist's microguide to environmental data", "Things are still quite complicated because we have more data available than before but it is often difficult to interpret and to use with journalistic tools" [9].

The revelations of data journalism often exceed people's ability to perceive them in textual format, so data-driven journalism (or data journalism) is developing alongside data visualization, and they have been shaping new genres and new sections in the media. Newsrooms have moved much farther along than primary data search and data processing. Algorithms can produce not just crude analytics for human journalists to consider but ready-to-publish textual and non-textual news products of a very specific nature.

The following story is already a milestone in the history of robo-journalism. On 17 March 2014, at 6:25 a.m., journalist and programmer of the *Los Angeles Times*, Ken Schwencke, was jolted awake by an earthquake. He rushed to his laptop, where he found an e-mail notification sent to him by an algorithm named Quakebot:

> ***L.A. Now: Ready for copyedit: Earthquake: 4.7 quake strikes near Westwood, California***
>
> *This is a robopost from your friendly earthquake robot. Please copyedit & publish the story. You can find the story at: [. . .]*
>
> *If the city referenced in the headline is relatively unknown, but the earthquake occurred close to another, larger city, please alter the headline and body text to put that information first.*

> *I am currently not smart enough to make these decisions on my own, and rely on the help of intelligent humans such as yourselves.*
>
> *Thanks! Quakebot* [10].

Schwencke loaded up the *Los Angeles Times'* content management system and found Quakebot's ready-to-publish report:

> *A shallow magnitude 4.7 earthquake was reported Monday morning five miles from Westwood, California, according to the U.S. Geological Survey. The temblor occurred at 6:25 a.m. Pacific time at a depth of 5.0 miles.*
>
> *According to the USGS, the epicenter was six miles from Beverly Hills, California, seven miles from Universal City, California, seven miles from Santa Monica, California and 348 miles from Sacramento, California. In the past 10 days, there have been no earthquakes magnitude 3.0 and greater centered nearby.*
>
> *This information comes from the USGS Earthquake Notification Service and this post was created by an algorithm written by the author.*
>
> *Read more about Southern California earthquakes* [11].

Schwencke thought it looked good, set it live and sent out a tweet [12].

By that time, Quakebot, which was written by Schwencke himself, had been around for two years. Quakebot pulled data (place, time, magnitude of earthquakes) from the U.S. Geological Survey's Earthquake Notification Service. Then, the robot compared these data to previous earthquakes in this area and defined "the historic significance" of the event. The data were then inserted into suitable sentence patterns, and the news report was ready. The robot uploaded it into the content management system, and sent a note to the editor.

Thus, the *LAT* became the first media outlet to report on the earthquake—eight minutes after the tremor struck and much earlier than any human journalists managed to report [13]. The intern robo-journalist had outrun its bio-colleagues.

This earthquake report was far from being worthy of a Pulitzer. However, it allowed an editor to publish a news story minutes after the event happened. Needless to say, earthquakes are hot news items in LA, and a fast response time adds significant value to reporting on them.

Another example of data-driven journalism also relates to the *Los Angeles Times*. The paper's section of criminal chronicles called "The Homicide Report" [13] has been run by an algorithm since 2007.

As soon as a coroner adds information on a violent death into the database, the robot draws all available data, places it on the map, categorizes it by race, gender, cause of death, police involvement, etc., and publishes a report online. Then, if deemed necessary, a human journalist can gather more info and write an extended criminal news story.

Robot participation has significantly changed traditional criminal coverage. In the past, a journalist would cover only newsworthy crimes, which were crimes with the highest resonance potential. Now, the robot covers absolutely all incidents involving death. The robot also allows data to be observed by categories of race, gender, and neighborhood, and visualized on a scalable and "timeable" county map, which creates secondary content of high importance that otherwise would have been missed in human reporting. The map of crimes composed by the robots has additional value, for instance, for the real estate market. Thus, algorithms add value to news by extracting insights from data analysis that are often unnoticeable to human reporters.

It is worth adding that the *Los Angeles Times'* criminal reporting robot covers a territory with a population of 10 million. This is comparable to the population of Sweden or Portugal. Certainly, a bio-journalist is not capable of doing instant statistical calculations on such a scale and quickly (instantly) putting it into synchronic and diachronic contexts within a format that is immediately accessible and easily comprehensible for readers.

Data mining or, more generally, data journalism is particularly in demand for coverage of areas with significant amounts of data—finances, weather, crime, and sports. As has already been stated, data mining can detect highly valuable content that simply cannot be traced by human reporters.

In addition, the internet of things is opening new horizons for data journalism. For example, digitalization of sports creates a completely new type of sports reporting. As stated by Steven Levy back in 2012 in his article for *Wired* [14], sports leagues have covered every inch of the field and each player with cameras and tags. Computers gather all possible data that one can imagine, such as ball speed, altitude, throwing distance, and the positions of players or even their hands—all made possible through telemetry. A well-trained robot can spot that a pitcher threw his last fastball a little weaker or that a batter leaned left before hitting a winning run. Is this information important? It is, but a human reporter would not notice it. Old-style sports journalism cannot do it, just like the old form of criminal reporting cannot produce an interactive map of the density of murders in an area.

2.2. Topic Selection

The principles of big data analysis and correlation analysis allow algorithms to make quick and precise decisions about what is newsworthy right now by assessing the interests of the audience. These interests manifest themselves in measurable ways: likes, shares, reposts, time spent, etc.

The amount of data that allow measuring human reactions to content will increase, particularly after biometric measurements of human reactions come into play. Eye tracking technologies are already used to understand the nuances of human attention. Algorithms are potentially able to decide what is of interest to humans, with any possible correlation between social-demographic categories and topics of interest.

Robots in newsrooms have already started doing this. Canada’s *Globe and Mail* uses an algorithm that traces readers’ preferences. Editors still decide which story to develop, but a robot suggests topics “that already have a proven track record with readers online”. As publisher Phillip Crawley described in an interview with Canadian Press, “Instinct of an experienced editor . . . can’t ever be substituted, but when you’ve got data which constantly feeds and gives you great clarity, there will be great surprises” [15].

The algorithm is able to complete more sophisticated tasks than just tracking what readers like, read, post, discuss (how many of them, for how long, etc.). Having enough data, it is only logical to take a step toward summarizing and analyzing readers’ motives and desires. The next step will be assignment planning. It has been reported that the *Globe and Mail* “also recently hired a technology expert with a PhD in artificial intelligence to design a ‘predictive modeling’ platform to help determine which stories will interest readers and drive engagement, such as sharing on social media” [15].

This case shows that robots can potentially substitute not just reporters but also editors. At this point, again, algorithms cannot make final decisions about what is interesting for the audience. Also problematic is the idea of thematic planning based on readers’ former preferences. However, a sufficient amount of data about readers’ behavior along with a strong enough predictive model and an instant and constant connection to all relevant sources of potential news can make such editor’s associates very potent and resourceful.

At the end of the day, the editor guesses, but the robot knows. When newsrooms have less data about readers, the editor with their guesswork has all the advantages. With more data about readers, the robot can take over the job.

And yet another consideration is worth mentioning. The analysis of readers’ (human) behavior is at only the beginning of its long and potentially infinite path. Both the quantity of data about the audience and the quality of the processing and cross-processing of these data will accelerate and grow endlessly; whereas, in contrast, the human potential for doing the editor’s job has already fully revealed itself. We are finishing when they are starting.

2.3. Commentary Moderation

Another area for algorithms' application in the media is fostering healthy conversation online, or commentary moderation. Audience engagement is an important asset in the media business. However, free access to commenting often provokes trolls and spammers. That is why, after falling in love with user-generated content at the end of the 2000s, many newsrooms shut down comment sections in the mid-2010s. Moderation took too much resource to maintain.

It looked strange when news outlets fenced themselves off from the audience. Many explanations were given by the media. See, for example, "*Huffington Post* to ban anonymous comments" [16] or "Online comments are being phased out" [17]. *Popular Science* even made a minor sensation in the industry in 2013 with its manifesto "Why We're Shutting Off Our Comments. Starting today, PopularScience.com will no longer accept comments on new articles. Here's why" [18].

In the late 2010s, a solution seems to have been found. Advanced newsrooms have started applying algorithms to regulate commentaries. For example, the *Washington Post* and the *New York Times* in collaboration with the Mozilla Foundation founded the Coral Project, a project that produces open-source software to maintain online communications within and outside newsrooms. Using algorithms, editors and reporters can survey readers, moderate comment sections, and engage the audience in many other ways [19].

As the Coral Project promotes itself, our plugin architecture gives publishers incredible flexibility to select the features that make sense for your community—either across your site, or on a single article.

For Commenters

- *Identify journalists in the conversation;*
- *Mute annoying voices;*
- *Manage your history;*
- *Sort by most replied/liked/newest;*
- *Follow and link to single discussions;*
- *See new comment alerts instantly;*
- *Manage separate identities on each site.*

For Moderators

- *Feature the best comments and filter out the worst;*
- *Use AI-assisted moderation to identify problems quickly;*
- *See detailed commenter histories, and take bulk actions;*
- *Integrate with industry-leading spam and abuse technologies;*
- *Moderate faster via Slack integration, keyboard shortcuts, and more.*

For Publishers

- *Own and manage all of your users' data;*
- *Connect to your existing login system;*
- *Reward subscribers/donors with badges—or restrict commenting only to them;*
- *Make your comments match your site design;*
- *Translate into any language your audience speaks* [20].

Such a detailed description is given here to show how powerful and helpful this tool can be. Most interestingly, as with any new medium, it does not just improve old functions (moderation, which would have been impossible for humans to execute on this scale), but also introduce new functions, such as organizing user-generated content for reporters' further use or for capitalizing on community involvement.

The *Washington Post* was the first news organization that integrated the Coral Project software called Talk with Modbot, the *Post's* own "AI-powered comment moderation technology". As they described the technology, "Talk's moderation panel serves up statistics to help moderators understand a commenter's contribution history at a glance. Then using ModBot, the system can remove comments that violate *Post* policies, approve comments that don't, and provide analytics for moderators about the tenor of a conversation" [21].

The *New York Times* uses another algorithm (they also directly call it AI in their reports) "to host better conversations". It is reported that,

> *[NYT] turned to Perspective, a free tool developed by Jigsaw and Google that uses machine learning to make it easier to host good conversations online. Perspective finds patterns in data to spot abusive language or online harassment, and it scores comments based on the perceived impact they might have on a conversation. Publishers can use that information to give real-time feedback to commenters and help human moderators sort comments more quickly. And that means news organizations and content publishers can grow their comment sections instead of turning them off, and can provide their readers with a better, more engaging community forum* [22].

Before using algorithms for moderation, the *Times* struggled to maintain healthy conversations in comments and was only able to enable comments on about 10 percent of articles. After engaging machine-learning algorithms to help human moderators, "The *New York Times* was able to triple the number of articles on which they offer comments, and now have comments enabled for all top stories on their homepage" [22].

Can algorithms that help moderate commentaries be called a "narrow artificial intelligence?" Maybe not yet. The idea is rather to show in which direction the development of the media is moving. Even if comment-moderating algorithms do not deserve to be called AI yet, they execute a job, in one part of which they have already greatly exceeded humans, and in another part they can already be compared to humans. Namely, moderating algorithms have exceeded in comments tracking in terms of speed and volume. No human can compare to a machine in this. But more intriguing is the ability of algorithms to assess commentaries.

The assessment of other people's tone and connotation is considered a human privilege and prerogative. However, algorithms are good not just at screening flagged words and expressions, but they are already able to perform semantic analysis. Even if they do not "understand" the essence of the offensive or hate speech concepts, they can use human reactions to comments as a tool of assessment. Then, machine learning comes into play.

2.4. Text Writing

Quakebot of the *Los Angeles Times* is able to collect and compare data, but the robot also composes ready-to-publish text reports. These reports, of course, are very simple since the robot just uses a set of templates. A criminal reporting robot does not write at all; it filters and categorizes data, puts them on a map, and so on. Analyses of such cases by media critics suggest that robots save time for human journalists, so that humans can do other, truly creative jobs.

The cases of financial and sports robo-journalism are much more complicated. In these fields, robots do not save time for humans; they take over the job entirely.

An algorithm called Wordsmith is one of the most hired and probably the most voluminous news-writing platforms. The algorithm developed by the hi-tech company Automated Insights can analyze data and put them into a coherent narrative with adjusted styles. According to Automated Insights' website, "Wordsmith is the natural language generation platform that transforms your data into insightful narrative" [23].

One of Wordsmith's employers, Associated Press, uses the platform to produce earnings reports. Each quarter, companies release earnings and news agencies inform their subscribers about companies' ups and downs. The speed, accuracy, and analyticity of reporting are important, because subscribers make business decisions based on these reports. So, news agencies make their business on these earnings recaps.

Associated Press had been able to produce only 300 earnings recaps per quarter before it hired Wordsmith. Thousands of potentially important company earnings used to be left unreported. The other problem related to the workload of reporters—earning recaps were "the quarterly bane of the existence of many business reporters". As *New York Magazine's* Kevin Roose put it, corporate earnings were "a miserable early-morning task that consisted of pulling numbers off a press release, copying them into a pre-written outline, affixing a headline, and publishing as quickly as possible so that traders would know whether to buy or sell" [24].

Media automation has solved both problems. The use of Wordsmith increased the Associated Press coverage of corporate earnings over tenfold. Now, the robot writes 4400 earnings stories per quarter. For the robot, it takes seconds to pull numbers from the earnings report, to compare them to previous data from the same company and data of competitors, to make a set of simple conclusions, to compose a smooth narrative structure, and to publish the story. Unlike its organic colleagues, the robot does not complain about how the job is boring and meaningless.

Here is a news story written by the Wordsmith algorithm and published by the Associated Press:

> *Apple tops Street 1Q forecasts*
>
> *Apple posts 1Q profit, results beat Wall Street forecasts*
>
> *AP. 27 January 2015 4:39 p.m.*
>
> *CUPERTINO, Calif. (AP) _ Apple Inc. (AAPL) on Tuesday reported fiscal first-quarter net income of $18.02 billion. The Cupertino, California-based company said it had profit of $3.06 per share. The results surpassed Wall Street expectations. The average estimate of analysts surveyed by Zacks Investment Research was for earnings of $2.60 per share. The maker of iPhones, iPads and other products posted revenue of $74.6 billion in the period, also exceeding Street forecasts. Analysts expected $67.38 billion, according to Zacks. For the current quarter ending in March, Apple said it expects revenue in the range of $52 billion to $55 billion. Analysts surveyed by Zacks had expected revenue of $53.65 billion. Apple shares have declined 1 percent since the beginning of the year, while the Standard & Poor's 500 index has declined slightly more than 1 percent. In the final minutes of trading on Tuesday, shares hit $109.14, an increase of 39 percent in the last 12 months* [25].

Such recaps can be compiled within less than a second. The robot seizes the facts of the earnings report, makes necessary market synchronic and diachronic comparisons, and generates a rather profound and reasonably coherent text. The news is full of data but made in a meager style. Still, it is a decent narrative. All in all, a financial report does not require stylish decorations.

While experts discuss the perspectives of algorithms' applications in the media, the scale of algorithms' applications is already beyond what one might imagine. In 2013, Wordsmith wrote 300 million stories. According to Lance Ulanoff from *Mashable*, this is more than all the major media companies combined [26]. In 2014, Wordsmith wrote 1 billion stories [27]. In 2016, it wrote 1.5 billion stories [28]. This is probably more than the work of all human journalists combined.

Wordsmith is not the only cyber reporter hired by the media. In the early 2010s, a company called Narrative Science developed StatsMonkey, a writing platform that generated baseball game recaps from applicable data such as players' activities, game scores, and win probability. Here is a fragment of a children's baseball league game report written by StatsMonkey,

> *Friona fell 10–8 to Boys Ranch in five innings on Monday at Friona despite racking up seven hits and eight runs. Friona was led by a flawless day at the dish by Hunter Sundre, who went 2–2 against Boys Ranch pitching. Sundre singled in the third inning and tripled in the fourth inning . . . Friona piled up the steals, swiping eight bags in all . . .* [14].

StatsMonkey's unique trait was that it used baseball slang. That was not its only benefit. Children's games' results could be input by parents into a special iPhone app during the game. StatsMonkey processed statistics and generated texts almost immediately. The fans, the little baseball players' Moms and Dads, received a recap of the match even before the little players finished shaking hands on the field. It goes without saying that such recaps, no matter their writing style, were much more important to these fans than a Super Cup report.

In 2011, StatsMonkey wrote 400,000 reports for the children's league. In 2012, it wrote 1.5 million [14]. For reference, that year, there were 35,000 journalists in the USA [29]. They likely would not be willing to cover Little League games, regardless of how much money they were offered to do so. That is another aspect of robot journalism—algorithms can cover topics that are skipped by human reporters for not being "newsworthy" These topics still find highly loyal readers.

After StatsMonkey, Narrative Science developed an "advanced natural language generation platform" called Quill. Quill analyzes structured data and automatically generates "comprehensive narrative reporting and personalized customer communications" [30] that can be used in the media but also in all sorts of financial market communications.

Narrative Science rented out Quill's writing skills to financial customers such as T. Rowe Price, Credit Suisse, and USAA. As a company representative said, "We do 10- to 15-page documents for some financial clients". However, Quill also wrote for Forbes [31]. So, as is reflected in the title of an article about it, "Robot Journalist Finds New Work on Wall Street" [32]. In fact, Quill the robo-journalist in part repeated the professional trajectory of many human financial journalists. After succeeding in financial analysis and reporting, some writers transition to the investment industry to write narrative-based and comprehensive financial reports for investors, partners, and clients. For Quill, as well as for human journalists, working for investment companies is probably more rewarding than for media organizations.

There are also other companies producing natural language generation software. *Columbia Journalism Review* listed 11 providers of automated journalism solutions in different countries in 2016, stating that,

> *Thereof, five are based in Germany (AX Semantics; Text-On; 2txt NLG; Retresco; Textomatic), two in the United States (Narrative Science; Automated Insights) and France (Syllabs; Labsense), and one each in the United Kingdom (Arria) and China (Tencent). The field is growing quickly: the review is not even published yet, and we can already add another provider from Russia (Yandex) to the list* [25].

In the UK, local newspapers have become involved in the automation project Urbs Media, which is endorsed by a 706,000 euro grant from Google. It aims to create 30,000 localized news reports every month. Urbs Media chose a natural language generation platform developed by Arria "to provide the AI backbone of its service" [33].

Leading media companies such as the Associated Press, *Forbes*, the *New York Times*, the *Los Angeles Times*, and ProPublica have started to automate news content [25]. Many news organizations have also begun developing their own, in-house news writing platform. In fact, an amount of news coverage generated by robots has already been huge. Even though many reports on robo-journalism suggest human reporters not worry about job security and look for ways of collaboration with robots, there are a lot of reasons for worries. The robots already beat humans in quantity; it may occur that assumed superiority of humans in quality is very much overestimated.

Not all cases of automated journalism implement artificial intelligence. However, narrow AI is undoubtedly involved at least in some projects, particularly ones in which robots already squeeze out humans.

It is also logical that news organizations integrate all their automated efforts within newsrooms. Data mining (data journalism), topic selection, commentary moderation (community development), and, finally, text writing are not separate tasks in newsrooms. All these processes are interrelated. Being organized around an intelligent platform, all these tasks not only can be better executed separately, but they also heighten the level of organizational coherence and integration. Some news organizations have already started to build AI-related intelligent news platforms of this next level.

Thus, the Chinese Xinhua News Agency, "has introduced the 'Media Brain' platform to integrate cloud computing, the Internet of Things, AI and more into news production, with potential applications 'from finding leads, to news gathering, editing, distribution and finally feedback analysis'" [34]. As Emily Bell, a professor at the Columbia Journalism School, commented on Twitter, "There are already elements of this in quite a few newsrooms but this is the first announcement (I've seen) of a large news org rearranging itself around AI" [34].

3. Turing Test in Journalism: Passed

The most frequent question in discussions about the future of robot-human competitions in journalism is, "Are robots capable of writing better than humans?"

In other words, robots have already surpassed human reporters in data journalism and also in speed and in scale of news coverage. But can they beat humans in writing style?

This question implies two assumptions that are questionable themselves. First of all, do humans write well? What humans? All of them? Second, does the robot need to write better than whom? Salinger and Tolstoy?

In fact, the question about robots' capability to excel beyond humans in writing implies conducting of a sort of Turing test in journalism. The journalistic Turing test would differ from the classical one, of course—it is not interactive. In the classical Turing test, a human asks a robot (not knowing that it is a robot) and is able to challenge an interlocutor with tricky questions in order to reveal if it is a human or an algorithm. In the journalistic Turing test, there is no interactive aspect; it is just the perception of a completed story as written by a human or a robot.

Such journalistic Turing tests have already been conducted.

In May 2015, Scott Horsley, an NPR White House correspondent and former business journalist, boldly challenged Wordsmith. "We wanted to know: How would NPR's best stack up against the machine?" NPR wrote [35]. As NPR is a radio station, then a bio-journalist working for them should be very well trained in fast reporting. According to the rules of the competition, both competitors waited for Denny's, the restaurant chain, to come out with an earnings report. Horsley had an advantage, as he was a Denny's regular. He even had a regular waitress, Genevieve, who knew his favorite order: Moons Over My Hammy. It did not help; although it depends on how the results are judged.

The robot completed the task in two minutes. It took Scott Horsley a bit more than seven minutes to finish. NPR published both news pieces to offer readers a sort of journalistic Turing test.

> ***Story #1***
>
> *Denny's Corporation on Monday reported first-quarter profit of 8.5 million dollars. The Spartanburg, South Carolina-based company said it had profit of 10 cents per share. The results beat Wall Street expectations. The average estimate of four analysts surveyed by Zacks Investment Research was for earnings of 9 cents per share. The restaurant operator posted revenue of $120.2 million in the period, also beating Street forecasts. Three analysts surveyed by Zacks expected $117.1 million. Denny's shares have risen nearly 6 percent since the beginning of the year. In the final minutes of trading on Monday, shares hit $10.90, a climb of 61 percent in the last 12 months.*
>
> ***Story #2***
>
> *Denny's Corporation notched a grand slam of its own in the first quarter, earning a better-than-expected ten cents a share, as restaurant sales jumped by more than 7-percent. Operating revenues topped $120 million. Adjusted net income jumped 36 percent to $8.7 million. Denny's is one of the nation's largest full-service restaurant chains. The growth in sales suggests consumers are opening their pocketbooks for pancakes, eggs, and hash browns. Earnings were also helped by lower costs for raw materials. Denny's results were also helped by the re-opening of the high-volume location inside the Las Vegas Casino Royale restaurant. After sales grew faster than expected in the first three months of the year, managers raised their sales forecast for the remainder of 2015* [35].

The first story was, of course, written by the robot. The figure density is higher and the style is drier. In the meantime, Horsley added a bit of unnecessary information into his version of the financial report, for instance, with this sentence: "the growth in sales suggests consumers are opening their pocketbooks for pancakes, eggs and hash browns".

Technically, the robot's vocabulary is larger, as it includes the entire national language's vocabulary. That is more than 1 million English words. But the robot has to use the most relevant, most conventional, i.e., most frequent words, and that eventually dries up its style. Moreover, the robot's vocabulary is limited by industrial specialization. To give an example, a robot would never use cooking or sports vocabulary in a financial report.

Humans are the opposite. An educated native English speaker boasts a vocabulary of around only 100,000 words. But a human writer is not limited by word relevancy or frequency, and has the freedom to use the rarest and most colorful words, which broadens context and brings vividness. Moreover, an original style of writing, often "deviating" from rational necessity, is what really makes a human a writer. A good writer can use the wrong wording deliberately, which is completely impossible for a robot writer (even though if it were programmed so, the wrong wording by a robot would not be deliberate in this case; the same goes for a bad human writer—they can use unsuitable words, but not intentionally). Robots simply do not "feel" a need for originality to complete a financial report.

"But that could change", NPR suggests [35]. If the owner supplies Wordsmith with more versatile NPR stories and modifies the algorithm a little, this kind of redesign could broaden and diversify Wordsmith's vocabulary. Such things are modifiable. Robo-journalists still will not get the necessity for originality, but they will be able to simulate stylish diversity at a level at which the artificiality of such style coloring will not be noticeable.

Here we are approaching the idea that a sufficiently large number of variances can compensate for writing algorithms' lack of "senses", at least at the level of routine consumer perception. The adding of a "random word generator" with specific stylistic instructions can make the product (text) as colorful as a person would do (not to mention that there are not many demands on human journalists regarding the colorfulness of the style).

So, who won the competition? The robot wrote faster and in a more business-like fashion. Scott Horsley was more "human-like" (which makes sense) but slower. The target audience of this writing consists of people in the financial industry. Is the lyrical digression about wallets and pancakes valuable to them? As long as readers are humans and not other robots, it might be.

The result of the contest can most likely be recognized as a tie. However, two minutes against seven for writing the story can be significant margin for radio and financial reporting, where the time of response matters.

Interestingly, the Turing test assesses the human quality of style, but not human productivity, or human accuracy. It implies that style is the most difficult constituent of the human speech (and social) faculty for robots to overcome. The ability of algorithms to overcome other constituents of human speech that are essential for the media has raised no questions.

Another journalistic Turing test, though in a humorous form and for the purposes of entertainment, was offered by the *New York Times*. The *Times* composed a quiz that allows readers to guess whether a human or an algorithm wrote a story.

1. "A shallow magnitude 4.7 earthquake was reported Monday morning five miles from Westwood, California, according to the U.S. Geological Survey. The temblor occurred at 6:25 a.m. Pacific time at a depth of 5.0 miles".

   ○ Human
   ✓ Computer

   (This excerpt of an initial report about a March 2014 earthquake was written by an algorithm.—*The commentary was opened after passing the quiz; I answered correctly.—A.M.*)

2. “Apple’s holiday earnings for 2014 were record shattering. The company earned an $18 billion profit on $74.6 billion in revenue. That profit was more than any company had ever earned in history”.

   ✓ Human
   ○ Computer

   (This was an excerpt from an article from Business Insider.—*The commentary was opened after passing the quiz; I answered incorrectly.—A.M.*)

3. “When I in dreams behold thy fairest shade Whose shade in dreams doth wake the sleeping morn The daytime shadow of my love betray’d Lends hideous night to dreaming’s faded form”.

   ○ Human
   ✓ Computer

   (This is an excerpt of a poem written by a poetry app.—*The commentary was opened after passing the quiz; I answered incorrectly.—A.M.*)

4. “Benner had a good game at the plate for Hamilton A’s-Forcini. Benner went 2–3, drove in one and scored one run. Benner singled in the third inning and doubled in the fifth inning”.

   ○ Human
   ✓ Computer

   (This was a sample report done by Quill, a Narrative Science product.—*The commentary was opened after passing the quiz; I answered correctly.—A.M.*)

5. “Kitty couldn’t fall asleep for a long time. Her nerves were strained as two tight strings, and even a glass of hot wine, that Vronsky made her drink, did not help her. Lying in bed she kept going over and over that monstrous scene at the meadow”.

   ○ Human
   ✓ Computer

   (The Russian novel “True Love” was written by a computer in St. Petersburg in 72 h.—*The commentary was opened after passing the quiz; I answered incorrectly, being deceived, of course, by the names of Kitty and Vronsky, characters from Tolstoy’s “Anna Karenina”.—A.M.*) . . . [36].

As for some brief reflections on the quiz: even knowing (or because of knowing) the advancement and distinctiveness of robo-journalism, even knowing (or because of knowing) that robots are already producing financial and sport reports, I was not able to differentiate human writing from algorithmic writing confidently. Furthermore, some prejudices about robots’ incapability to compose poetry (although they can; and it is already a well-known fact) forced me to make a mistake and assume that a piece of poetry was written by a human.

That said, at this level of media consumption, robo-writers have already passed the journalistic Turing test. The naivety of this test’s conditions reproduces a real situation of media consumption, so it only strengthens its adequacy and proves the adequacy of its result.

Academics staged a competition between “a horse and a steam-engine”, too. Christer Clerwall, a Media and Communications professor from Karlstad, Sweden, asked 46 students to read two reports [37]. One was written by a robot and another by a human. The human news story was shortened to the size of the robot one. The robot news story was slightly edited by a human, so that its headline, lead, and first paragraphs looked similar to what is usually done by an editor. Students were asked to evaluate the stories based on certain criteria such as objectivity, trust, accuracy, boringness, interest, clarity, pleasure to read, usefulness, coherence, etc.

The results showed that one of the news stories led in certain parameters, and the other one excelled in others. The human story led in such categories as "well-written", "pleasant to read", etc. The robot news story won the categories "objectivity", "clear description", "accuracy", etc. So humans and robots tied again.

But the most important thing the Swedish study revealed is that the differences between the average text of a human writer and a cyber-journalist are insignificant. The distinction between human-written and robot-written texts is approximately the same as between texts written by different humans. They both can be accepted by an editor.

This is a crucial argument for assessing the future and even the current state of robo-journalism. Cyber-skeptics have frequently argued that robots cannot write better than humans. But this is the wrong way to approach the issue. As professor Clerwall tells *Wired*, "Maybe it doesn't have to be better—how about 'a good enough story'?" [38].

In the *New York Times's* article "If an Algorithm Wrote This, How Would You Even Know?" Shelley Podolny states that, "Robots make pretty decent writers" [27]. Indeed, even when measuring them by human standards, we can see that they write, if not better, then at least not worse than humans. At the very least, human readers cannot confidently distinguish robot and human writing.

The question about the ability of AI to replace humans in journalism usually comes down to the question "Are algorithms able to write better than humans?" This line of thinking, in fact, is incorrect. To be a journalist, humans do not need to write better than humans (again, what humans? Dostoyevsky? Twain?). Humans only have to write good enough. The same goes for algorithms. To be hired in the media, robots do not have to write better than humans—they have to write good enough. And they do.

4. Counterarguments about "Robots' Incapability"

The other doubt about AI in journalism relates to machines' inability to simulate human creative talents.

In the media, the "creativity counterargument" can be represented by doubts in the ability of an algorithm (1) to invent/discover; (2) to distinguish beauty/originality in writing.

Let us examine these doubts.

4.1. A Robot Cannot Invent/Discover

Yes, certainly, it is hard to imagine a robot exclaiming "Eureka!" Serendipity is a human gift. A human can come upon accidental inventions or discoveries for no reason (like when an apple falls on one's head). At the same time, human inventors are most often capable of recognizing an invention, even accidentally made. Thus, the invention/discovery is characterized by a strange combination of preparedness and impossibility to confidently prearrange. No computational or linear process can lead to it. It is impossible to calculate or code an invention/discovery.

Regarding journalism, invention/discovery can relate to the idea of creating a topic or developing texture. This mandatory part of the job of a journalist, creativity, seems to be impossible for a robot to reproduce.

That is why skeptics would say that robots will not be able, for example, to "smell" a potential sensation in a sequence of same-type events, as a human editor easily does. Moreover, a robot will not be able to decide to overblow a story, as human editors do by intentionally picking an event from a seemingly indistinguishable mass of similar ones.

However, what if, in turn, robots can do something that humans cannot; some other kinds of novelty, if not invention?

This potential novelty, the new knowledge that humans cannot obtain and robots can, relates to the cross-analysis of big data and correlations. The ability to see correlations behind big data is incomprehensible to humans but can be considered as something substituting, for robots, the human faculty of invention/discovery.

We humans value causation over correlation, probably because of our lack of computational skills. Many correlations revealed through the analysis of big data seem strange to us. Tyler Vigen in his book *Spurious Correlations* [39] presents more than 30,000 correlations extracted from big data that seemingly make no any sense. For example, US spending on science, space, and technology has correlated with the number of suicides by hanging, strangulation and suffocation over a decade, 1999–2009, with astonishing precision. In the same decade, the number of people who drowned by falling into a pool closely correlated with the number of films in which Nicolas Cage appeared. The divorce rate in Maine closely correlated with the per capita consumption of margarine in 2000–2009, and so on.

Always looking for causality but not being able to find it behind correlation, we reject any sense when considering spurious correlations. However, some of these correlations may make someone think twice. For example, the total revenue generated by arcades almost precisely correlates with the number of computer science doctorates awarded in the US in 2000–2009. What if there is something behind such coincidences?

Properly trained algorithms can examine correlations between two variables, but also between three, 33, or 3000 variables. The range of correlations can be enormous and, in fact, unlimited. Anything could correlate to anything through anything else. Unlike Eureka or serendipitous human moments, it is calculable. We cannot even imagine the limits of this intellectual operation, as the size of databases and the processing speed of machines has been growing constantly. But if correlations in massive amounts of data are detected, it may mean something. For humans, it makes sense only if correlations are reduced to causation. For algorithms, causation is not a motive for processing information. AI can process data without searching for causality. This is a type of intellectual motivation that is very distinctive from that of humans.

Those weird but sustained correlations that are so easily detected by algorithms look like magic to us; but it also can be new knowledge or even a new type of knowledge, which is always a sort of magic for those to whom it is incomprehensible.

This reasoning shows that robots have something at their disposal regarding the novelty of knowledge, too. How will they manage this ability to learn new things? Hypothetically, the ability to detect correlations leads to the potential for learning and revealing everything, possibly in a way that we may not understand. It depends on the amount of big data and the processing speed.

Through such a perspective, a robot's lack of inventive/discovery talents seems like an increasingly unimportant disadvantage. Robots have opportunities to find fantastic correlations that are sometimes of great practical importance for marketing, politics, or media. The world is full of them. They work without an explanation through causality, and bio-journalists are simply unable to see them.

What if an algorithm's ability to identify correlations compensates or even outdoes the human skills of invention/discovery? Facts derived from big data may be as interesting and irrational as outcomes of human creativity. Another relevant consideration: we are already aware of the potential of human creativity, while big data processing and correlation detection are just at the beginning of their potentially endless path.

4.2. Robots Do Not Understand Beauty or Originality

Yes, it is true; robots do not aim to write in a beautiful manner. Even if they had such a goal, what would be defined as "beauty"? What is that?

However, even if it is impossible to calculate beauty, it is possible to calculate human reactions to it. Humans themselves can serve robots as a new type of servomechanism—beauty-meters.

It is quite possible to detect correlations between texts, headlines, or even certain expressions, on the one hand, and human reactions to them, such as likes, shares, comments, and click-troughs, on the other hand. Also, the "size" of big data matters. The more texts and headlines with human reactions to them an algorithm will obtain, the more precise its "understanding" of the human perception of beauty will be. Even today, robots are able to identify the attractiveness of headlines, topics, keywords, etc., by observing people's reactions. Editors guess, robots know.

Robots' capacity to read human reactions will only grow. An algorithm developed by Facebook is already customizing newsfeeds according to users' reactions, from which personal preferences can be calculated. With the help of biometrics, robots will be able to analyze human physiological reactions to certain semantic and idiomatic expressions, epithets, syntax structures, and visual images.

If algorithms do not have their own senses, humans can serve as receptors that convert sensory reflexes to computable signals. As tools and mechanisms once were the extensions of humans, humans can now be good extensions for machines, allowing machines to enhance their faculties and reach out beyond their "natural" limitations.

Maybe it is time to revisit the idea of who serves whom. Marshall McLuhan, in his *Playboy* interview, foresaw that "man becomes the sex organs of the machine world just as the bee is of the plant world, permitting it to reproduce and constantly evolve to higher forms" [40]. In *Understanding Media*, he wrote that, "By continuously embracing technologies, we relate ourselves to them as servomechanisms. That is why we must, to use them at all, serve these objects, these extensions of ourselves, as gods or minor religions. An Indian is the servo-mechanism of his canoe, as the cowboy of his horse or the executive of his clock" [41] (p. 46).

By operating beauty-meters that rely on the measuring of human reactions, algorithms will be able to automatically produce more attractive texts and headlines without understanding the concept of beauty (or originality, or style).

In other words, for every argument about what robots cannot do, there is a more convincing argument for what robots can do instead. In this competition of capabilities, robots and humans also end up in a tie.

The competition has just begun, but it is a tie already. Humans are an old team, while the younger robot team is just making its debut.

5. Roadmap for Artificial Intelligence in the Media—And Beyond

Considering possible AI accomplishments that will allow it to bypass its lack of creativity, it is possible to outline the future developments of artificial intelligence in the media. This will lie in three main interrelated areas: data processing, data accumulation, and understanding human reactions.

(1) Data processing. Algorithms are used to develop ways to manage big data. Their ability to find correlations will in some way replace human creativity. The best human minds are working on it now—they are working on facilitating algorithms for the best possible performance. These brilliant human minds—coders, developers, and engineers from the real and symbolic Silicon Valley—do not care about saving journalism. They aim to implement their innovations and tools without any reservations and often even without any moral consideration. For the replacement of humans by robots, humans (and the smartest humans) will be responsible, not robots.

"If there is a free press, journalists are no longer in charge of it. Engineers who rarely think about journalism or cultural impact or democratic responsibility are making decisions every day that shape how news is created and disseminated", said Emily Bell, professor at the Columbia Journalism School, in her speech with a title that speaks for itself, "Silicon Valley and Journalism: Make up or Break up?" [42].

(2) Data accumulation. As everything now leaves its footprint on the Internet, the database of all texts and all audiences' reactions will be able to be collected sooner or later. By monitoring, gathering, and analyzing all journalistic texts and people's reactions to them, algorithms will be able to calculate which texts with which parameters earn more likes, reads, reposts, and comments.

(3) Understanding of human reactions. Learning of human reactions is one of the most important tasks for artificial intelligence in its efforts to convince us that it can replace us. Understanding human reactions has already become a crucial factor in social media and marketing, led by algorithms. The same goes for the media. By now, algorithms' ability to read human reactions comes down to the analysis of likes, reposts, time spent with content, etc. But this will change.

Once robots get access to human non-verbal reactions and body language, they will be able to calculate inexplicit reactions instantly. For instance, if someone reads a story about Trump and has certain somatic reactions, technologies already are capable of reading some of them. Webcams can scan the movement of pupils, microphones can hear an increase in breathing frequency, while detectors added to touchscreens could sense the heartbeat or perspiration, and so on.

Even based on relatively "rational" human reactions (likes, reposts, time spent) algorithms can comprehend the audience's preferences better than human editors. After biometrics is incorporated, algorithms' understanding of the audience's preferences will reach the forensic preciseness of a polygraph.

Interestingly, due to the development of editors' skills required from algorithms in the media, artificial intelligence with biometrics detectors will become a very inquisitive and enormously powerful polygraph for all of society—a global polygraph for the global village, another version of the Orwellian Big Brother, but introduced into the Huxleyan *Brave New World*. Biometrics development will be not determined by any final cause; it rather will be market-driven. Meaning it will proceed and succeed.

These three aspects of algorithmic development are important not only within the media industry. They also will possibly pave the way for the real—general, or strong, AI to come.

Narrow AI in the media has already arrived. When it integrates all newsrooms tasks, including the editor's job of assignment allotting, it will approach the idea of goal setting. Smart robo-editors able to pick topics, set tasks, and measure human reactions will obtain power over the entire production cycle in the media. There is little need to say how important this is in terms of agenda setting and ruling over people. Society has already faced similar problems with Facebook algorithms, although this is just a forerunner of future problems.

Ultimately, an intelligence with the ability to set goals for itself is already no longer an intelligence; it is rather a being with its own will.

6. Conclusions. Will They Replace Journalists? Yes

In 2012, Kristian Hammond, co-founder of Narrative Science, predicted that algorithms will write 90% of media content by 2030. As *Wired* quoted him, "In 20 years, there will be no area in which Narrative Science doesn't write stories" [14].

The author of that article in *Wired*, Steven Levy, wrote, "As the computers get more accomplished and have access to more and more data, their limitations as storytellers will fall away. It might take a while, but eventually even a story like this one could be produced without, well, me".

Interestingly, Hammond also predicted that a computer would write a story worth a Pulitzer Prize "in five years", which meant in 2017. This did not happen. However, this symbolic act of awarding of the Pulitzer to a robo-journalist will no doubt eventually happen, as it happened in 2016 when the Nobel Prize in literature was awarded to Bob Dylan with an obvious intention to mark some new trend in literature, or rather to mark the death of old literature. This will happen to journalism, too.

Concluding the review of possible AI developments regarding its use in the media, it can be said that human journalists are in a qualitative and quantitative competition with cyber-colleagues. This competition is not at the beginning, as the general public would think. It is moving to the end. In the quantitative contest, bio-journalists have already lost. They are set to lose in the qualitative competition in 5–7 years.

It is also interesting that in the period of transition from a predominantly organic form of journalism to a predominantly cybernetic form of journalism, it will be humans, not robots, who drive the process forward. First, developers, coders, and engineers just do their job well and without any visible reason to stop. Second, it will be editors who will hire robots to write, to moderate comments, and to select topics. In fact, it will be editors who kill the profession for humans and turn it to a function performed by robots.

The reason is simple—the economics of the media business. Newsrooms have to produce as much content as possible in order to increase traffic, views, click-through rates, etc. After switching from the “portioned” production of periodicals/TV/radio to the “streaming” production of the internet, the media have to run more and more stories. Online means non-stop. It is motion for motion’s sake. Media theorist Dean Starkman called this effect the “hamsterization of journalism” [43]. The hamsterization of journalism reduces time spent by the journalist on each story in order to produce more stories: “do more with less”.

Let us imagine that a good article, which means good journalism, can attract thousands of readers. But what if a thousand news stories written over the same time period were able to attract just a hundred readers each? Actually, when traffic is king, editors do not need the best journalists; they need fast journalists . . . Whom will the editor choose—a capricious, talented (or not so much) journalist with increasing salary demands and three stories per week or a flawless algorithm with decreasing maintenance costs that can produce three stories per minute?

The Associated Press buys the Wordsmith service not because the algorithm writes better than humans. The reason is that the algorithm writes both more and faster. Debates about the quality of the writing are not relevant. Robots will conquer newsrooms not for belletristic reasons, but for economic ones.

If humans still preserve jobs in the media, it will happen not because of economic reasons, but rather because of the social need to utilize people. It happens to many industries: the preservation of jobs becomes more important than increasing efficiency. Socialism is beating capitalism in this way. This is the only considerable reason for people to stay in the media, and it is beyond the context of competition with algorithms.

Thus, robots’ advent in the media is unstoppable. Under these conditions, the most beneficial strategy for newsrooms is to be among the first at the beginning of robotization and the last at the end of it.

For now, the editorial use of algorithms could be an interesting PR strategy that is attractive for both audiences and investors. But when algorithms fill the market, the rare human voice will be in demand amid the chorus of robots.

In this sense, as strange as it seems, human journalism will be particularly valued at the final stages of robotization of the media as a distinct flavor. Moreover, editorial human errors will be particularly valued and attractive, and human-made media will capitalize on errors. That is going to happen at least until robots learn to simulate human errors, too (in order to better substitute humans).

If Wordsmith published 1.5 billion stories in 2016, a part of these stories increased the physical amount of content. The other part, however, already was to replace human writing. This is easy to see as exemplified by Associated Press. By the time AP hired Wordsmith, human reporters were writing 300 earnings recaps per quarter. Wordsmith writes 4400. This gives us a probable quantitative pattern: robo-journalists produce 10-times the total volume of content and drive out that former humans’ share of 300 recaps.

The market is going to demand more and more. Nothing can stop robots from writing as much as they are required to, since the only limit for them is the amount of content that people can read. Even this limit will be removed, once the readers are also robots.

Funding: This research received no external funding.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Russell, S.J.; Norvig, P. *Artificial Intelligence: A Modern Approach*, 3rd ed.; Prentice Hall: Upper Saddle River, NJ, USA, 2009.
2. McLuhan, M. *Understanding Media: The Extensions of Man*; McGraw Hill: New York, NY, USA, 1964.
3. De Chardin, P.T. *The Phenomenon of Man*; Harper: New York, NY, USA, 1959.
4. Military Applications of Artificial Intelligence. *Bulletin of the Atomic Scientist*, April 2018. Available online: https://thebulletin.org/military-applications-artificial-intelligence (accessed on 30 June 2018).

5. Kerner, I. What the Sex Robots Will Teach Us. *CNN*, 13 March 2018. Available online: https://www.cnn.com/2016/12/01/health/robot-sex-future-technosexuality/index.html (accessed on 30 June 2018).
6. Owsianik, J. State of Sex Robots: These Are the Companies Developing Robotic Lovers. *Futureofsex.net*, 16 November 2017. Available online: https://futureofsex.net/robots/state-sex-robots-companies-developing-robotic-lovers/ (accessed on 30 June 2018).
7. Lieberman, H. In Defense of Sex Robots. *Quartz*, 2 March 2018. Available online: https://qz.com/1215360/in-defense-of-sex-robots/ (accessed on 30 June 2018).
8. Jeremy, B. How The New York Times Uses Software to Recognize Members of Congress. *Times Open*, 6 June 2018. Available online: https://open.nytimes.com/how-the-new-york-times-uses-software-to-recognize-members-of-congress-29b46dd426c7 (accessed on 30 June 2018).
9. Bouchart, M. A Data Journalist's Microguide to Environmental Data. *Data Journalism Blog*, 15 January 2018. Available online: http://www.datajournalismblog.com/2018/01/15/data-journalists-microguide-environmental-data/ (accessed on 30 June 2018).
10. Meyer, R. How a California Earthquake Becomes the News: An Extremely Precise Timeline. *The Atlantic*, 19 March 2014. Available online: https://www.theatlantic.com/technology/archive/2014/03/how-a-california-earthquake-becomes-the-news-an-extremely-precise-timeline/284506/ (accessed on 30 June 2018).
11. Oremus, W. First Earthquake Report Written by a Robot. Screenshot. Source: The First News Report on the L.A. Earthquake Was Written by a Robot. *Slate*, 17 March 2014. Available online: http://www.slate.com/blogs/future_tense/2014/03/17/quakebot_los_angeles_times_robot_journalist_writes_article_on_la_earthquake.html (accessed on 30 June 2018).
12. Schwencke, K. How to Break News While You Sleep. *Source*, 24 March 2014. Available online: https://source.opennews.org/articles/how-break-news-while-you-sleep/ (accessed on 30 June 2018).
13. The Homicide Report. *The Los Angeles Time*. Available online: http://homicide.latimes.com/ (accessed on 30 June 2018).
14. Levy, S. Can an Algorithm Write a Better News Story than a Human Reporter? *Wired*, 24 April 2012. Available online: https://www.wired.com/2012/04/can-an-algorithm-write-a-better-news-story-than-a-human-reporter/ (accessed on 30 June 2018).
15. Globe and Mail to Tap into Online Data to Help Reshape Daily Newspaper. *Canadian Press*, 6 September 2017. Available online: http://j-source.ca/article/globe-mail-tap-online-data-help-reshape-daily-newspaper/ (accessed on 30 June 2018).
16. Landers, E. Huffington Post to Ban Anonymous Comments. *CNN*, 22 August 2013. Available online: https://edition.cnn.com/2013/08/22/tech/web/huffington-post-anonymous-comments/index.html (accessed on 30 June 2018).
17. Gross, D. Online Comments Are Being Phased out. *CNN*, 21 November 2014. Available online: https://edition.cnn.com/2014/11/21/tech/web/online-comment-sections/) (accessed on 30 June 2018).
18. LaBarre, S. Why We're Shutting off Our Comments. Starting Today, PopularScience.com Will No Longer Accept Comments on New Articles. Here's Why. *Popular Science*, 24 September 2013. Available online: https://www.popsci.com/science/article/2013-09/why-were-shutting-our-comments (accessed on 30 June 2018).
19. Erickson, T. Will Comment Sections Fade away, or Be Revived by New Technologies? *MediaShift*, 19 January 2018. Available online: http://mediashift.org/2018/01/will-comment-sections-fade-away-revived-new-technologies/ (accessed on 30 June 2018).
20. The Coral Project. Available online: https://coralproject.net/products/talk.html (accessed on 30 June 2018).
21. The Washington Post Launches Talk Commenting Platform. *WashPost PR Blog*, 6 September 2017. Available online: https://www.washingtonpost.com/pr/wp/2017/09/06/the-washington-post-launches-talk-commenting-platform/ (accessed on 30 June 2018).
22. New York Times: Using AI to Host Better Conversations. *Blog Google*. Available online: https://blog.google/topics/machine-learning/new-york-times-using-ai-host-better-conversations (accessed on 30 June 2018).
23. Automated Insights. Available online: https://automatedinsights.com/wordsmith (accessed on 30 June 2018).
24. Roose, K. Robots Are Invading the News Business, and It's Great for Journalists. *New York Magazine*, 11 July 2014. Available online: http://nymag.com/daily/intelligencer/2014/07/why-robot-journalism-is-great-for-journalists.html (accessed on 30 June 2018).

25. Graefe, A. Guide to Automated Journalism. *Columbia Journalism Review*, 7 January 2016. Available online: https://www.cjr.org/tow_center_reports/guide_to_automated_journalism.php (accessed on 30 June 2018).
26. Ulanoff, L. Need to Write 5 Million Stories a Week? Robot Reporters to the Rescue. *Mashable*, 2 July 2014. Available online: https://mashable.com/2014/07/01/robot-reporters-add-data-to-the-five-ws/#Hmwz.0hssgqi (accessed on 30 June 2018).
27. Podolny, S. If an Algorithm Wrote This, How Would You Even Know? *The New York Times*, 7 March 2015. Available online: https://www.nytimes.com/2015/03/08/opinion/sunday/if-an-algorithm-wrote-this-how-would-you-even-know.html (accessed on 30 June 2018).
28. Allen, R. The AI Entrepreneur's Moral Dilemma. *Machine Learning in Practice Blog*, 12 July 2017. Available online: https://medium.com/machine-learning-in-practice/the-ai-entrepreneurs-moral-dilemma-12b988f18cd0 (accessed on 30 June 2018).
29. Up against the Paywall. *The Economist*, 19 November 2015. Available online: https://www.economist.com/business/2015/11/19/up-against-the-paywall (accessed on 30 June 2018).
30. Quill. Narrative Science Web Site. Available online: https://narrativescience.com/Products (accessed on 30 June 2018).
31. Morozov, E. A Robot Stole My Pulitzer! How Automated Journalism and Loss of Reading Privacy May Hurt Civil Discourse. *Slate*, 19 March 2012. Available online: http://www.slate.com/articles/technology/future_tense/2012/03/narrative_science_robot_journalists_customized_news_and_the_danger_to_civil_discourse_.single.html (accessed on 22 July 2018).
32. Simonite, T. Robot Journalist Finds New Work on Wall Street. *Technology Review*, 9 January 2015. Available online: https://www.technologyreview.com/s/533976/robot-journalist-finds-new-work-on-wall-street/ (accessed on 30 June 2018).
33. Marr, B. Another Example of How Artificial Intelligence Will Transform News and Journalism. *Forbes*, 18 July 2017. Available online: https://www.forbes.com/sites/bernardmarr/2017/07/18/how-a-uk-press-agency-will-use-artificial-intelligence-to-write-thousands-of-news-stories-every-week/#5019bd1474db (accessed on 30 June 2018).
34. Schmidt, C. China's News Agency Is Reinventing Itself with AI. *NiemanLab*, 10 January 2018. Available online: http://www.niemanlab.org/2018/01/chinas-news-agency-is-reinventing-itself-with-ai/ (accessed on 30 June 2018).
35. Smith, S.V. An NPR Reporter Raced a Machine to Write a News Story. Who Won? *NPR*, 20 May 2015. Available online: https://www.npr.org/sections/money/2015/05/20/406484294/an-npr-reporter-raced-a-machine-to-write-a-news-story-who-won (accessed on 30 June 2018).
36. Did a Human or a Computer Write This? *The New York Times*, 7 March 2015. Available online: https://www.nytimes.com/interactive/2015/03/08/opinion/sunday/algorithm-human-quiz.html?smid=pl-share&_r=0 (accessed on 30 June 2018).
37. Clerwall, C. Enter the Robot Journalist. Users' perceptions of automated content. *J. Pract.* **2014**, *8*, 519–531.
38. Clark, L. Robots Have Mastered News Writing. Goodbye Journalism. *Wired*, 6 March 2014. Available online: http://www.wired.co.uk/article/robots-writing-news (accessed on 30 June 2018).
39. Vigen, T. *Spurious Correlations*; Hachette Books: New York, NY, USA, 2015.
40. McLuhan, M. The Playboy Interview. *Playboy Magazine*, March 1969.
41. McLuhan, M. *Understanding Media: The Extensions of Man*; The MIT Press: Cambridge, MA, USA, 1994.
42. RISJ Admin. *Silicon Valley and Journalism: Make up or Break up*; The Reuters Institute for Study of Journalism, University of Oxford: Oxford, UK, 2014; Available online: http://reutersinstitute.politics.ox.ac.uk/risj-review/silicon-valley-and-journalism-make-or-break (accessed on 30 June 2018).
43. Starkman, D. The Hamster Wheel. Why Running as Fast as We Can Is Getting Us Nowhere. *Columbia Journalism Review*, September/October 2010. Available online: https://archives.cjr.org/cover_story/the_hamster_wheel.php?page=all (accessed on 30 June 2018).

information

Opinion

Artificial Intelligence Hits the Barrier of Meaning †

Melanie Mitchell [1,2]

[1] Department of Computer Science, Portland State University, Portland, OR 97207-0751, USA; mm@pdx.edu
[2] Santa Fe Institute, Santa Fe, NM 87501, USA

† This article is reprinted from Mitchell M. Artificial Intelligence Hits the Barrier of Meaning, *New York Times*, 5 November 2018.

Received: 20 January 2019; Accepted: 2 February 2019; Published: 5 February 2019

Abstract: Today's AI systems sorely lack the essence of human intelligence: Understanding the situations we experience, being able to grasp their meaning. The lack of humanlike understanding in machines is underscored by recent studies demonstrating lack of robustness of state-of-the-art deep-learning systems. Deeper networks and larger datasets alone are not likely to unlock AI's "barrier of meaning"; instead the field will need to embrace its original roots as an interdisciplinary science of intelligence.

Keywords: deep neural networks; meaning; understanding

You've probably heard that we're in the midst of an AI revolution. We're told that machine intelligence is progressing at an astounding rate, powered by "deep learning" algorithms that use vast amounts of data to train complicated programs known as "neural networks." Today's AI programs can recognize faces and transcribe spoken sentences. We have programs that can spot subtle financial fraud, find relevant web pages in response to ambiguous queries, map the best driving route to most any destination, beat human grandmasters at chess and Go, and translate between hundreds of languages. What's more, we've been promised that self-driving cars, automated cancer diagnoses, housecleaning robots, and even automated scientific discovery are on the verge of becoming mainstream. Facebook founder Mark Zuckerberg recently declared that, over the next five to ten years, the company will push its AI to "get better than human level at all of the primary human senses: vision, hearing, language, general cognition" [1]. Shane Legg, Chief Scientist of Google's DeepMind group, predicted that "human-level AI will be passed in the mid-2020s" [2].

As someone who has worked in AI for decades, I've witnessed the failure of similar predictions of imminent human-level AI, and I'm certain these latest forecasts will fall short as well. The challenge of creating humanlike intelligence in machines remains vastly underestimated. Today's AI systems sorely lack the essence of human intelligence: Understanding the situations we experience, being able to grasp their meaning. The mathematician and philosopher Gian-Carlo Rota famously asked, "I wonder whether or when AI will ever crash the barrier of meaning." To me, this is still the most important question concerning AI.

The lack of humanlike understanding in machines is underscored by recent cracks that have appeared in the foundations of modern AI. While today's programs are much more impressive than the systems we had twenty or thirty years ago, a series of research studies have shown that deep-learning systems can be unreliable in decidedly unhuman-like ways.

I'll give a few examples. "The bareheaded man needed a hat" is transcribed by my phone's speech-recognition program as "The bear headed man needed a hat." Google Translate renders "I put the pig in the pen" into French as "Je mets le cochon dans le stylo" (mistranslating "pen" in the sense of a writing instrument). Programs that "read" documents and answer questions about them can easily be fooled into giving wrong answers when short, irrelevant snippets of text are appended to the document [3]. Similarly, programs that recognize faces and objects, lauded as a major triumph of deep learning, can fail dramatically when their input is modified even in modest ways by certain types of lighting, image filtering, and other alterations that do not affect humans' recognition abilities in the slightest [4]. One recent study showed that adding small amounts of noise to a face image can seriously harm the performance of state-of-the-art face-recognition programs [5]. Another study, humorously called "The Elephant in the Room", showed that inserting a small image of an out-of-place object, such as an elephant, in the corner of a living-room image strangely caused deep-learning vision programs to suddenly misclassify other objects in the image [6]. Furthermore, programs that have learned to play a particular video or board game at a "superhuman" level are completely lost when the game they have learned is slightly modified (e.g., the background color on a video-game screen is changed, or the virtual "paddle" for hitting "balls" changes position) [7,8].

These are only a few examples demonstrating that the best AI programs can be unreliable when faced with situations that differ—even in a small degree—from what they have been trained on. The errors made by such systems range from harmless and humorous to potentially disastrous: Imagine, for example, an airport security system that won't let you board your flight because your face is confused with that of a criminal, or a self-driving car that, due to unusual lighting conditions, fails to notice that you are about to cross the street.

Even more worrisome are recent demonstrations of the vulnerability of AI systems to so-called "adversarial examples", in which a malevolent hacker can make specific changes to images, sound waves, or text documents, which, while imperceptible or irrelevant to humans, will cause a program to make potentially catastrophic errors. The possibility of such attacks has been demonstrated in nearly every application domain of AI, including computer vision, medical image processing, speech recognition, and language processing. Numerous studies have demonstrated the ease with which hackers could, in principle, fool face- and object-recognition systems with specific minuscule changes to images [9], put inconspicuous stickers on a stop sign to make a self-driving car's vision system mistake it as a yield sign [10], or modify an audio signal so that it sounds like background music to a human but instructs a Siri or Alexa system to perform a hidden command [11].

These potential vulnerabilities illustrate the ways in which current progress in AI is stymied by the barrier of meaning. Anyone who works with AI systems knows that behind the facade of humanlike visual abilities, linguistic fluency, and game-playing prowess, these programs do not—in any humanlike way—understand the inputs they process or the outputs they produce. The lack of such understanding renders these programs susceptible to unexpected errors and undetectable attacks.

What would be required to surmount this barrier, to give machines the ability to more deeply understand the situations that they face, rather than have them rely on shallow features? To find the answer, we need to look to the study of human cognition. Our own understanding of the situations we encounter is grounded in broad intuitive "commonsense knowledge" about how the world works, and about the goals, motivations, and likely behavior of other living creatures, particularly other humans. Additionally, our understanding of the world relies on our core abilities to *generalize* what we know, to form abstract concepts, and to make analogies—in short, to flexibly adapt our concepts to new situations. Researchers have been experimenting for decades with methods for imbuing AI systems with intuitive common sense and robust humanlike generalization abilities, but as yet there has been little progress in this very difficult endeavor.

AI programs that lack common sense and other key aspects of human understanding are increasingly being deployed for real-world applications. While some people are worried about "superintelligent" AI, the most dangerous aspect of AI systems is that we will trust them too much and give them too much autonomy, while not being fully aware of their limitations. As AI researcher Pedro Domingos noted, "People worry that computers will get too smart and take over the world, but the real problem is that they're too stupid and they've already taken over the world" [12].

The race to commercialize AI has put enormous pressure on researchers to produce systems that work "well enough" on narrow tasks. But ultimately, the goal of developing trustworthy AI will require a deeper investigation into our own remarkable abilities, and new insights into the cognitive mechanisms we ourselves use to reliably and robustly understand the world. Unlocking AI's barrier of meaning will likely require a step backward for the field, away from ever bigger networks and more massive data collections, and back to the field's original roots as an interdisciplinary science studying the most challenging of scientific problems: The nature of intelligence.

Conflicts of Interest: The author declares no conflict of interest.

References

1. McCracken, H. Inside Mark Zuckerberg's Bold Plan for the Future of Facebook (2015). *Zuckerberg Transcripts*. 16 November 2015. Available online: https://dc.uwm.edu/zuckerberg_files_transcripts/210 (accessed on 3 February 2019).
2. Despres, J. Scenario: Shane Legg. Available online: http://future.wikia.com/wiki/Scenario:_Shane_Legg (accessed on 4 December 2018).
3. Jia, R.; Liang, P. Adversarial Examples for Evaluating Reading Comprehension Systems. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 1–7 September 2017.
4. Hendrycks, D.; Dietterich, T.G. Benchmarking neural network robustness to common corruptions and surface variations. *arXiv* **2018**, arXiv:1807:01697.
5. Goswami, G.; Ratha, N.; Agarwal, A.; Singh, R.; Vatsa, M. Unravelling robustness of deep learning based face recognition against adversarial attacks. In Proceedings of the American Association for Artificial Intelligence (AAAI 2018), New Orleans, LA, USA, 2–7 February 2018; pp. 6829–6836.
6. Rosenfeld, A.; Zemel, R.; Tsotsos, J.K. The elephant in the room. *arXiv* **2018**, arXiv:1808.03305.
7. Rusu, A.A.; Rabinowitz, N.C.; Desjardins, G.; Soyer, H.; Kirkpatrick, J.; Kavukcuoglu, K.; Pascanu, R.; Hadsel, R. Progressive neural networks. *arXiv* **2016**, arXiv:1606.04671.
8. Kansky, K.; Silver, T.; Mély, D.A.; Eldawy, M.; Lázaro-Gredilla, M.; Lou, X.; Dorfman, N.; Sidor, S.; Phoenix, S.; George, D. Schema Networks: Zero-Shot Transfer With a Generative Causal Model of Intuitive Physics. In Proceedings of the International Conference on Machine Learning (2017), Stockholm, Sweden, 10–15 July 2018; pp. 1809–1818.
9. Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; Fergus, R. Intriguing Properties of Neural Networks. In Proceedings of the International Conference on Learning Representations (2014), Banff, MB, Canada, 14–16 April 2014.
10. Eykholt, K.; Evtimov, I.; Fernandes, E.; Li, B.; Rahmati, A.; Xiao, C.; Prakash, A.; Kohno, T.; Song, D. Robust Physical-World Attacks on Deep Learning Visual Classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018), Salt Lake City, UT, USA, 18–22 June 2018; pp. 1625–1634.
11. Carlini, N.; Wagner, D. Audio Adversarial Examples: Targeted Attacks on Speech-To-Text. In Proceedings of the First Deep Learning and Security Workshop (2018), San Francisco, CA, USA, 24 May 2018.
12. Domingos, P. *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*; Basic Books: New York, NY, USA, 2015.

information

Article

Technological Singularity: What Do We Really Know?

Alexey Potapov [1,2]

[1] Department of Computer Photonics and Videomatics, ITMO University, 197101 Saint Petersburg, Russia; potapov@aideus.com

[2] SingularityNET, 1083 HN Amsterdam, The Netherlands; alexey@singularitynet.io

Received: 27 February 2018; Accepted: 3 April 2018; Published: 8 April 2018

Abstract: The concept of the technological singularity is frequently reified. Futurist forecasts inferred from this imprecise reification are then criticized, and the reified ideas are incorporated in the core concept. In this paper, I try to disentangle the facts related to the technological singularity from more speculative beliefs about the possibility of creating artificial general intelligence. I use the theory of metasystem transitions and the concept of universal evolution to analyze some misconceptions about the technological singularity. While it may be neither purely technological, nor truly singular, we can predict that the next transition will take place, and that the emerged metasystem will demonstrate exponential growth in complexity with a doubling time of less than half a year, exceeding the complexity of the existing cybernetic systems in few decades.

Keywords: technological singularity; metasystem transitions; artificial general intelligence

1. Introduction

Technological progress is visibly accelerating. There are many exponential trends in addition to the commonly known Moore's law, including the increase in the number of computers connected to the Internet, or of the amount of data acquired by neuroimaging technologies. Moreover, the more the technologies develop, the steeper the exponential growth becomes. So, the overall progress seems not simply exponential, but hyper-exponential, asymptotically going to infinity within a finite time period, namely, within few decades (see, e.g., [1]). A point at which a function is not defined is called a singularity in mathematics. By analogy, a hypothetical point at which technological progress becomes unbounded is called a technological singularity (Singularity).

The idea of Singularity excites many people (including myself), who naturally try to speculate about its possible implications. Some speculations may seem to go too far (e.g., [2]) with strong but ungrounded statements and predictions. However, this should not devalue the idea itself, and criticism of such ungrounded claims should not be considered an argument against the Singularity correctly understood.

One of the most frequently encountered misconceptions consists in identifying the Singularity with the creation of artificial (super) intelligence. Even the Wikipedia article [3] on the technological singularity starts with the statement that the invention of artificial superintelligence will be the cause of Singularity (although it later indicates that Vernor Vinge, who popularized the notion of Singularity [4], wrote about other ways to Singularity).

While artificial superintelligence is one possible path to the Singularity, it is not the only one. Many critics identify the Singularity with this one possibility (e.g., [5]), which itself is controversial, at least, if it is taken to imply that this superintelligence must be achieved by modern computers. The Singularity will not necessarily come about through the creation of Strong AI with digital computers.

The responsibility for such misconceptions lies with the adepts of the Singularity themselves, since they not infrequently assert all their beliefs and desires simultaneously, assuming that they can thus reinforce each other. At the same time, their critics seek the weakest points assuming that refuting them will render the whole concept invalid. They are also usually biased by their desires which can be simply expressed as "Singularity is impossible because we do not want humans to disappear."

Here, I will try to disentangle the grounded claims from the personal beliefs and desires, and underline what we can really say about Singularity. In particular, I will give a definition of Singularity in terms of metasystem transitions as an objective phenomenon without referring to any particular technology.

2. What Do We Know

2.1. Metasystem Transitions

Although the theory of metasystem transitions is rarely mentioned in connection with the concept of Singularity (e.g., [6] mentions it specifically in the connection with the emergence of the Global Brain), it is an essential scientific foundation of this concept. This theory, proposed by Valentin Turchin in [7], was originally based on the study of the evolution of cybernetic (control) systems in the evolution of nervous systems. Evolution of these systems takes place through a sequence of metasystem transitions, each of which consists of a creation of a higher-level control system that chooses between states or different exemplars of already existing lower-level control system.

Let us consider a cybernetic system that has an internal state, for example, a spatial location. Control of this location by effectors is defined as a "motion," which itself initially is uncontrollable. Control of the motion is defined as an "irritability," which is enabled by the development of sensors. In time this leads to the development of control of the irritability as a simple reflex system, a coordinated but rigid reaction of effectors to certain patterns of sensory input. Control of multiple simple reflexes leads to the development of a more complex/conditional reflex (association). Control of associations is defined as "thought." And the control of thoughts (as defined in this cybernetic perspective) creates a culture (see Table 1).

Table 1. Stages of evolution [7].

Chemical era	Chemical forms of life
	Motion
	Irritability
Cybernetic era	Neural network (simple reflex)
	Association (conditional reflex)
Era of mind	Thought
	Culture, cultural integration

In addition, Valentin Turchin considered some sequences of metasystem transitions within culture (especially in mathematics), and we can identify similar metasystem transitions in many other systems as well. For example, different levels of gene control emerged during biological evolution, while the Internet can be considered as a metasystem with respect to computers.

Some extensions of this theory exist (e.g., [8]), but we will not go into detail here.

2.2. Timeline

What is missing in the theory of metasystem transitions is a timeline. The concept of Singularity is usually justified by timelines that track some key events in evolution supplemented by some qualitative measure, for example, memory capacities. Although different authors choose different key events as indicators, curves of growing complexity or decreasing time intervals between paradigm shifts as measured by key events are consistent as shown by Ray Kurzweil with 15 lists of key events [1].

These findings, the details of which I will not reproduce here, suggest that metasystem transitions representing global or universal evolution (e.g., [9]) follow two regular patterns:

1. Systems with a certain level of control grow exponentially (at least, before the next metasystem transition) in their capacities or complexity.
2. The time before the next metasystem transition decreases geometrically and the growth rate increases geometrically from transition to transition.

The Singularity is thus the point at which these patterns cease to exist.

These regularities are quite well grounded in the empirical data. It is difficult to deny that, for example, both the number of neurons in nervous systems and the number of transistors in computers have been growing exponentially, and the doubling time of the latter is much shorter. What is still uncertain is the significance of these observations and conclusions that can be drawn from them.

3. Predictions

We will distinguish two types of predictions about Singularity, namely, the extrapolation of its timeline and qualitative depiction of its possible scenarios (see, e.g., [1,4]).

3.1. Timeline Extrapolation

Extrapolation is probabilistic induction from past trends. If we do not use additional information, then the simplest extrapolation is the most probable one. Here, the simplest extrapolation suggests an accelerating sequence of metasystem transitions, such that complexity will grow to infinity in a finite period of time. True Singularity is this imaginary point, the date of which is somewhat uncertain, but most evidence suggests (see, e.g., [1]) that it is not more than few decades away.

There are many studies that try to predict when artificial general intelligence will emerge (e.g., [10–13]). However, I will not rely on these predictions here in order to not be involved in the controversy regarding the very possibility of thinking machines.

More interestingly, study [14] showed the synchronicity of different approaches to predicting long-term trends (including economic cycles, environmental and generational analysis besides purely technological trends) suggesting that there will be a technological surge in the 2040s, which might correspond to the Singularity if certain technologies become available at the time.

According to the current scientific picture of the world nothing can be truly infinite, so, such a "True Singularity" is thought to be physically impossible. Of course, our model of the world may change in future, but at this point the simple extrapolation of one curve does not provide sufficient grounds for changing it. Rather, it is more likely that this extrapolation will not continue indefinitely.

It is quite likely that the growth will decelerate at some point. This does not invalidate the concept of Singularity, because something that is not actually infinite can be close enough to infinite for any practical purpose. Thus, the real question is how high the complexity of cybernetic systems will grow.

The second simple extrapolation is an S-shaped curve. This posits that growth is exponential for a period of time, but that is slows down as it approaches some limit. There is reason to believe that this S-shaped curve is the usual pattern in the growth of complexity (as indicated, e.g., in [14]) in metasystem transitions, and this pattern is repeated in a fractal-like way on different time scales.

On a human time scale, the shape of the long-term curve is not critical. For humans, it does not even matter if the curve will saturate (or even fall down) at some point or will be unbounded. What does matter is when and how fast it will decelerate. There are no reasons to believe that such deceleration will be very rapid or abrupt.

The S-shaped curve is quite a conservative extrapolation, and we do not see signs that deceleration has already started. Thus, we still have not passed the inflection point, and after this point we will see slower but still rapid growth (excluding catastrophic scenarios). Thus, if the inflection point is a few decades off, this will be enough for an "Essential Singularity," after which the cybernetic systems at the cutting edge of universal evolution will become far more complex than the currently existing systems.

3.2. Possible Scenarios

No specific scenario can be considered to be a justified prediction, and thus criticism of a scenario cannot be used to criticize the general concept of Singularity. Does this mean that this concept cannot be used to make testable predictions, that is, that it does not satisfy Popper's criterion of falsifiability and thus is unscientific? Not precisely. We cannot say which specific metasystem transition will take place, but we can predict that some transition will most likely take place within a certain time range, and the emerged metasystem will demonstrate the exponential growth of its complexity with the doubling time less than half year exceeding the complexity of the existing cybernetic systems in few decades.

Nevertheless, we can try to assess which scenarios are relatively more or less probable. All scenarios are based on so-called Singularity technologies, that is, technologies that accelerate their own development, a phenomenon that usually depends on some form of superintelligence. For example, genetic engineering can help smarter humans to appear, who will then accelerate genetic research resulting in the emergence of even smarter humans.

Broad classes of possible Singularity technologies include bio-, nano-, info-, and maybe some other technologies, and their combinations. For example, one can talk about nanorobots populating human brains enhancing their capabilities, or about autonomous artificial general intelligence (AGI) optimizing itself and its own hardware. These technologies have different doubling times. For example, years are needed for genetically modified humans to be born and taught. Other forms of superintelligence can emerge much faster rendering the genetic modification route obsolete or supplementary especially taking the social factors associated with genetic engineering into account.

Of course, such an analysis is far from certain since it cannot take unknown future technologies into account, and does not consider all interactions between different technologies. It also does not take social, geopolitical, and economic factors into account, which might be necessary for predicting the future (see, e.g., [14]). Nevertheless, it can give us an educated guess about which technologies have a smaller doubling time and are most likely to lead to the next metasystem transition. However, my point here is that such predictions should not be used to criticize the concept of Singularity as such.

An additional source of prediction is the theory of metasystem transitions. For example, one might argue that the next metasystem will be "Control of Cultures." One can further argue that this is already happening in sense of humans interacting through the Internet with each other and with artificial agents, or that it will happen in a form of "Global Brain" (e.g., [6,15]). Although this looks like a logical consequence of the theory of metasystem transitions, this theory is not detailed enough to describe and predict the "hardware" of metasystems. For example, it says nothing about how nervous systems emerged as new hardware of cybernetic systems supplementing DNA. Similarly, it does not tell us what hardware is suitable for the level of culture and its consequent metalevels.

Formerly, the culture was "executed" by human brains augmented with external artefacts such as books. These cultural networks were similar to gene networks, but not to neurons. Will human brains still be the hardware for the cutting-edge of the universal evolution perhaps by being directly connected to each other? Or will computers become a metalevel system to control our culture through recommendations, and so forth? Will humans still be a part of the next metasystem, or will this system leave humans on the verge of universal evolution as an inefficient implementation? The theory of metasystem transitions does not provide definite answers to these questions. It simply says that most likely the next metasystem will be based on human culture, but does not say how exactly this will be implemented.

4. Misconceptions

4.1. True Singularity

The concept of "True Singularity" (taken not just as a simplified model, but as the reality) has a religious aspect, since it implies an emergence of an infinitely powerful god-like entity. As we have seen, there is minimal supporting evidence for this because it contradicts whole volumes of scientific data. Of course, hardly anyone believes in "True Singularity" in its ultimate form, but there is its soft version, when this entity remains finite, but occupies the whole Universe. This assumes that the speed of light limit can be overcome thanks to the development of "ontotechnologies" modifying the reality itself and its physical laws (why not if we can modify our genomes?). And the awakened Universe starts to communicate with other sentient universes within Multiverse. Although such ideas have some grounds in physical theories (regarding the possible place of our Universe in Multiverse [16,17]), they are just speculations.

Such ideas are fun, but should not be considered as real predictions. Vice versa, their implausibility should not be considered as a counter argument against Singularity per se.

However, it should be also noted that although many definitions of Singularity do not explicitly refer to the asymptotical technological progress, and the formal asymptotic limit of truly infinite progress cannot be achieved, mere exponential growth is not enough to achieve Singularity as discussed in [18].

4.2. Humans

What is the "Essential Singularity"? We can say that this is the point "in the history of the race beyond which human affairs, as we know them, could not continue" as it was formulated by Stan Ulam with the reference to his conversation with John von Neumann more than 60 years ago. However, this definition is far from definitive. On the one hand, many human affairs are quite different now from 200 years ago. On the other hand, some activities conducted by humans, such as science, could continue even without humans, at least as humans exist today. Will a genetically modified or augmented human still be a human? Is a human who uses a computer or a paper still a pure human? These are rhetorical questions.

We cannot define the "Essential Singularity" relative to humans who are permanently changing as a part of a larger metasystem. Maybe Singularity has already taken place in accordance with the 60-year-old definition. Whether we put it in such a way or not will not affect the reality. Of course, for humans, the fate of human life does matter, but this is difficult to predict. What we can say is that the universal evolution will continue, and metasystems transitions will take place leading to cybernetic systems of much greater complexity than that of a single human without tools.

4.3. We Have Choice

Universal evolution lasts for billions years. Its laws of metasystem transitions are not rigid, but they are objective. Evolution happens independently of our desires. Of course, humans are much more sentient beings than DNAs or single neurons. It seems we can choose. But can we choose to stop universal evolution? Hardly. Different people have different opinions on the question of how much we can shape evolution. Corporations and countries have their own interests. It is difficult to imagine that the development of all Singularity technologies (or, rather, all technologies) will be prohibited in all countries, and will not be performed by anyone in the world.

Similarly, people favor some scenarios over others, because they like them better. For example, transhumanists prefer to talk about brain uploading considering AGI as not too relevant or even as a threat. I do not try to assess the relative likelihood of these two scenarios, but simply compare them since they belong to the same paradigm (executing intelligence on computers). Modeling natural neurons on computers requires enormous computing resources. The computational resources needed for a human-level AI will be available much earlier than those required for executing an uploaded brain in real time. To reduce overhead, we need to precisely understand how to abstract away all the biochemical and physical details (e.g., 3D protein folding, which is extremely difficult task, but which is needed to model gene expression necessary for memory consolidation). Thus, we should already have a detailed model of (human) intelligence to do this.

AGI as a Singularity technology will also have a shorter doubling time, because knowledge of its design principles will enable easier self-optimization or extension with additional modules, sensory modalities, and so on. Thus, whether we want this or not, AGI will emerge earlier and evolve faster than brain uploading or whole brain emulation (if either of these is ever possible).

Governments, corporations, scientific societies and others can influence the speed of development of different technologies through financial support or restrictions, but this does not affect the inherent objective properties of these technologies, which play a major role in the pathway followed by universal evolution. Predictions of the plausibility of different scenarios should be based on the detailed comparison of the properties of different technologies and their possible mutual influence. We need to assess which technology is expected to appear earlier and develop faster and how this will influence the development of other technologies, and so forth This does not depend on our desires or preferences.

4.4. Artificial Superintelligence

As was mentioned, Singularity is frequently associated with the creation of artificial superintelligence (and even justified by it, for example, [19,20]). But this is also the source of criticism of the concept of Singularity itself [5].

The textbook example of a computer-based AI, which designs new faster computers and runs on them to design faster computers faster, and so forth, is just an illustration to the concept of "intelligence explosion." However, any other Singularity technology or a set of technologies can be substituted. For examples, humans use computers to conduct genetic research and to improve computers resulting in both smarter humans and faster computers accelerating both directions of research with positive feedback.

In this connection, I would like to make two claims:

- the concept of Singularity understood as a sequence of accelerating metasystem transitions does not depend on the idea of superhuman strong AI, and can be defended independently;
- the idea that superhuman general AI can be created in few decades is justified by evidence of the doubling times of different singularity technologies.

One might think that the second claim says the same thing as the above mentioned Wikipedia article [3]. But this is not really the case, because the causal relations are different. If one simply says that "the creation of AGI will lead to Singularity," then if we call the possibility of the creation of AGI into question, we will doubt even more about the coming Singularity.

On the contrary, we can substantiate the concept of Singularity independently of our assumptions about AGI, so even if we lean towards a negative answer to a quite controversial question about the possibility of AGI based on digital computers, this will not affect the plausibility of the concept of Singularity. Then, we can provide arguments that AGI (or, rather, non-human superintelligence) is most likely to emerge first. This is the independent (and weaker) claim, the possible fallacy of which does not affect the first claim.

Indeed, we saw that the concept of Singularity can be introduced independently of the concept of artificial superintelligence. The necessity for superintelligence to be artificial is an additional independent premise. Also, if superintelligence is not posited to require individual consciousness or strong integrity, then one can claim that such superintelligence has already been here for a long time, and is constantly becoming smarter and smarter (i.e., many tasks that were impossible for a mind armed only with pen and paper, have become doable with current technology), and for us there is no reason to believe that this process will suddenly terminate.

However, purely artificial superintelligence is also possible. There are no fundamental restrictions preventing this, especially, if we consider not only existing computers but also possible future computers, which can be based on other (possibly unknown now) physical processes. Even opponents of Strong AI such as John Searle and Roger Penrose addressed their criticism only to digital computers and not to all possible computing devices in principle. One can also add (e.g., [19]) that artificial superintelligence might not be necessary a Strong AI, but can be just a general AI, to which most criticism (based on subjective aspects of human intelligence like consciousness, qualia, etc.) is not applicable.

It can be regretted that progress does not enhance all the components of the human mind in equal degree. It is curious to note that the components that are the least affected are those that may more difficult to reproduce with computers. These are emotional intelligence, sense of humor, and so forth. I will not try to dispel these doubts here, but simply note that there are different opinions on this topic, and that theories of artificial creativity, curiosity, and fun exist (e.g., [21]). My main point here is that this is not a reason to deny the likelihood of progress per se. One can complain about the one-sidedness of this progress. One can also argue that it should not be called progress. However, this does not negate the fact of (hyper) exponential technological growth and, consequently, the concept of Singularity. It can also be posited that, for further progress, a strong AI (possessing all human qualities) might be not really necessary.

Thus, it is unscientific to claim that artificial general superintelligence in any form is strictly impossible, and disprove the concept of Singularity on this basis. However, we should not also claim that AGI, especially based on digital computers, would be an inevitable step towards Singularity.

Although personally I do believe that AGI can be created on the base of digital computers and it is most likely step towards Singularity due to the shortest doubling time, this is really a belief that might be false, so I neither want to defend it here, nor do I want its controversy to cast a shadow on the concept of Singularity.

5. Conclusions

One might choose to define a scenario with the creation of autonomous artificial superintelligence as Singularity, but others could define a Singularity as any scenario with the creation of any kind of superintelligence. Such discrepancy can be a source of controversy. Further, we can understand in different ways, what is “artificial” or what is “superintelligence”. We should not argue about definitions, but should be precise in what we claim.

Here, I have tried to disentangle two types of claims which can be defended independently, namely, the claims about the character of the technological progress (or, rather, universal evolution), and the claims about artificial intelligence.

I do not defend claims about AI here (although I found it necessary to mention some of them), and mainly focus on what we can say about Singularity, namely: some metasystem transition will most likely take place within a certain time range, and the emerged metasystem will demonstrate exponential growth of its complexity with the doubling time less than half year (implying that its hardware will not be limited to the biological components) exceeding the complexity of the existing cybernetic systems in few decades. Most likely the next metasystem will be based on exponential change in human culture (although this does not mean it cannot also involve an artificial superintelligence). One way or another, further metasystem transitions will take place, although their growth rate will start to decelerate at some point.

Will this future metasystem transition be a Singularity? It depends on definitions, and on a which scenario takes place which is difficult to predict. Thus, it is useless to argue about whether Singularity as a specific event will occur and (if yes) when. Strictly speaking, Singularity is a virtual time point at which the simplest extrapolation of the curve of growing complexity hits infinity, which will be never really achieved. However, all models in science describe the reality approximately, and they should not be criticized for this. Behind the concept of Singularity is the real phenomenon of accelerating universal evolution, which should not be discarded just because Singularity is a very simple predictive model which does not exhaust the phenomenon. All the criticism should be addressed to the use of the model independently of the specific scenario to which it is applied.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kurzweil, R. *The Singularity is Near*; Viking Books: New York, NY, USA, 2005; ISBN 978-0670033843.
2. Garis, H. How Will the Artilect War Start? In *The End of the Beginning: Life, Society and Economy on the Brink of the Singularity*; Goertzel, B., Goertzel, T., Eds.; Humanity+ Press: Leeds, UK, 2015; ISBN 0692457666.
3. Wikipedia: Technological Singularity. Available online: https://en.wikipedia.org/wiki/Technological_singularity (accessed on 22 February 2018).
4. Vinge, V. The Coming Technological Singularity: How to Survive in the Post-Human Era. In Proceedings of the Vision-21: Interdisciplinary Science and Engineering in the Era of Cyberspace, Cleveland, OH, USA, 30–31 March 1993; pp. 11–22.
5. Braga, A.; Logan, R.K. The Emperor of Strong AI is Naked: Limits to Artificial Intelligence. *Computers* **2018**. (submitted).
6. Heylighen, F. Return to Eden? Promises and Perils on the Road to Global Superintelligence. In *The End of the Beginning: Life, Society and Economy on the Brink of the Singularity*; Goertzel, B., Goertzel, T., Eds.; Humanity+ Press: Leeds, UK, 2015; ISBN 0692457666.
7. Turchin, V.F. *The Phenomenon of Science: A Cybernetic Approach to Human Evolution*, 1st ed.; Columbia University Press: New York, NY, USA, 1977; 348p, ISBN 978-0231039833.
8. Heylighen, F. (Meta)systems as Constraints on Variation: A classification and natural history of metasystem transitions. *World Futur. J. Gen. Evol.* **1995**, *45*, 59–85. [CrossRef]
9. Wikipedia: Universal Evolution. Available online: https://en.wikipedia.org/wiki/Universal_evolution (accessed on 22 February 2018).
10. Moravec, H. When will computer hardware match the human brain? *J. Transhumanism* **1998**, *1*, 10.
11. Bostrom, N. How Long Before Superintelligence? *Linguist. Philos. Investig.* **2006**, *5*, 11–30.
12. Baum, S.; Goertzel, B.; Goertzel, T. How long until human-level AI? Results from an expert assessment. *Technol. Forecast. Soc. Chang.* **2011**, *78*, 185–195. [CrossRef]
13. Grace, K.; Salvatier, J.; Dafoe, A.; Zhang, B.; Evans, O. When Will AI Exceed Human Performance? Evidence from AI Experts. *arXiv* **2017**, arXiv:1705.08807.
14. Goertzel, T.; Goertzel, B. (Eds.) Predicting the Age of Post-Human Intelligences. In *The End of the Beginning: Life, Society and Economy on the Brink of the Singularity*; Humanity+ Press: Leeds, UK, 2015; ISBN 0692457666.
15. Vidal, C. Distributing Cognition: From Local Brains to the Global Brain. In *The End of the Beginning: Life, Society and Economy on the Brink of the Singularity*; Goertzel, B., Goertzel, T., Eds.; Humanity+ Press: Leeds, UK, 2015; ISBN 0692457666.
16. Smolin, L. The status of cosmological natural selection. *arXiv* **2006**, arXiv:1705.08807.
17. Susskind, L. *The Cosmic Landscape: String Theory and the Illusion of Intelligent Design*, 1st ed.; Little, Brown: Boston, MA, USA, 2005; ISBN 978-0316155793.
18. Grey, A. The Singularity and the Methuselarity: Similarities and Differences. In *The End of the Beginning: Life, Society and Economy on the Brink of the Singularity*; Goertzel, B., Goertzel, T., Eds.; Humanity+ Press: Leeds, UK, 2015; ISBN 0692457666.
19. Muehlhauser, L.; Salamon, A. Intelligence Explosion: Evidence and Import. In *Singularity Hypotheses: A Scientific and Philosophical Assessment*; Eden, A.H., Moor, J.H., Soraker, J.H., Steinhart, E., Eds.; Springer: Berlin, Germany, 2012; pp. 15–42, ISBN 978-3642325601.

20. Loosemore, R.; Goertzel, B. Why an Intelligence Explosion is Probable. In *Singularity Hypotheses: A Scientific and Philosophical Assessment*; Eden, A.H., Moor, J.H., Soraker, J.H., Steinhart, E., Eds.; Springer: Berlin, Germany, 2012; pp. 83–98, ISBN 978-3642325601.
21. Schmidhuber, J. Formal Theory of Creativity, Fun, and Intrinsic Motivation (1990–2010). *IEEE Trans. Auton. Ment. Dev.* **2010**, *2*, 230–247. [CrossRef]

information

MDPI

Review

From Homo Sapiens to Robo Sapiens: The Evolution of Intelligence

Anat Ringel Raveh * and Boaz Tamir *

Faculty of interdisciplinary studies, S.T.S. Program, Bar-Ilan University, Ramat-Gan 5290002, Israel
* Correspondence: anatringel@gmail.com (A.R.R.); canjlm@actcom.co.il (B.T.)

Received: 30 October 2018; Accepted: 18 December 2018; Published: 21 December 2018

Abstract: In this paper, we present a review of recent developments in artificial intelligence (AI) towards the possibility of an artificial intelligence equal that of human intelligence. AI technology has always shown a stepwise increase in its capacity and complexity. The last step took place several years ago with the increased progress in deep neural network technology. Each such step goes hand in hand with our understanding of ourselves and our understanding of human cognition. Indeed, AI was always about the question of understanding human nature. AI percolates into our lives, changing our environment. We believe that the next few steps in AI technology, and in our understanding of human behavior, will bring about much more powerful machines that are flexible enough to resemble human behavior. In this context, there are two research fields: Artificial Social Intelligence (ASI) and General Artificial Intelligence (AGI). The authors also allude to one of the main challenges for AI, embodied cognition, and explain how it can be viewed as an opportunity for further progress in AI research.

Keywords: Artificial Intelligence (AI); Artificial General Intelligence (AGI); Artificial Social Intelligence (ASI); social sciences; singularity; complexity; embodied cognition; value alignment

1. Introduction

1.1. From Intelligence to Super-Intelligence

In this paper we present a review of recent developments in AI towards the possibility of an artificial intelligence equals that of human intelligence. So far AI technology has shown a stepwise increase in its capacity and in its complexity (Section 2). The last step took place several years ago, due to increased progress in deep neural network technology. Each such step goes hand in hand with our understanding of ourselves and our understanding of human cognition (Section 3). Indeed AI was always about the question of understanding human nature.

There is still a long way to go before we can talk about a singularity point. AI is still a weak technology, still too rigid, and too specific to become similar to human intelligence. However, we believe the next few steps in AI technology and in our understanding of human behavior, will bring about much more powerful machines, flexible enough to resemble human behavior. An important major research project in this context is Artificial Social Intelligence (ASI), which we shall shortly describe (Section 4). The second project is a new challenge which is known as Artificial General Intelligence (AGI) (Section 5). AGI brings about a new approach which is much more flexible and closer to human intelligence. It also suggests a model of consciousness, a new approach to the question of learning, a model of self-referential machines, etc.

One of the biggest challenges for AI is that of embodied cognition. If AI could surpass this hurdle, it will get even closer to true human intelligence. Embodied cognition was first presented as the main reason why AI is impossible. We propose to view embodied cognition as a new challenge to AI and not as an imposition (Section 6).

We end the discussion by demonstrating a way to overcome our fears of singularity, by the process of value alignment, which is expanded upon in Section 7.

1.2. The Human-Machine Eco-System

From ancient myths of inanimate objects coming alive to the creation of artificial intelligence, philosophers, scientists, writers, and artists have pondered the very nature and boundaries of humanity. Humans are fascinated by machines that can imitate us but also feel an existential discomfort around them—an uneasiness that stems from their ability to obscure the line between the living and the inanimate.

Claude Levi-Strauss [1] has examined how the individual process of constructing reality is related to how an entire society develops and maintains its worldview. He argued that the most common way in which both an individual and a community put together a structure of reality is through the use of binary categories. An individual makes sense of the world by organizing things in a series of dual oppositions such as dark/light, living/dead, feminine/masculine, emotion/logic, and so on, which lead to the community's development of more abstract concepts like, chaos/order, natural/unnatural, normal/abnormal, subjectivity/objectivity, and moral/immoral. Such a predetermined schema of reality provides the confidence people need to face the world and explore its boundaries.

As far as our relationship to thinking machines is concerned, it seems that the worldview we have developed for ourselves over time has become pessimistic as the pace of technology development increased. While in the past automata have entertained us mainly because they mimicked human behavior in an inaccurate and ridiculous way that revealed the fact that it was a trick, artificially intelligent machines today can successfully mimic an increasing number of the human's traits, such as natural human language and thought patterns. These traits have always separated us from all the other living creatures; and the fact that this primal distinction between human beings and technology is blurrier than it has ever been, mostly creates fear [2].

However, as Strate [3] explains, "At the very least what ought to be clear is that the physical universe and the biophysical environment are not entirely different and distinct from technology, but are part of a continuum." (p. 70). The view underling this argument is an ecological or systems view which "emphasize the interdependence and interactive relationships that exist, as all forms of life alter their conditions simply by their very presence, by their metabolism, for example, and through their reproduction." (p. 62).

Hence, instead of the dichotomous narrative of us—human beings—vs. them—super-intelligent machines, we should understand both ourselves and AI as parts of a complex and dynamic eco-system. While it might be that we are only a stepping stone on the path of the universe towards an even greater complexity, we have an important and special role; as argued by McLuhan [4], technologies and media "also depend upon us for their interplay and their evolution" (p. 57), and if we carefully examine their actions, we will see that there is no reason to fear.

2. State of the Art: Stepwise Incremental AI

In our view, the best way to describe the developmental process of AI so far is as a stepwise incremental progress, and using an analogy from physics, AI "percolates" into our lives. It could indeed make a phase transition into a higher level of complexity above our own, a process we will shortly discuss. But first we want to describe the ongoing process of stepwise incremental AI.

In January 2017 McKinsey [5] published a comprehensive report that maps the progress of artificial intelligence in a variety of areas. The five areas, described herein, are further broken down into other sub-tasks and sub-capabilities:

Sensory perception—"This includes visual perception, tactile sensing, and auditory sensing, and involves complex external perception through integrating and analyzing data from various sensors in the physical world." [5] (p. 34). In this area the performance level is median compared to humans.

Take machine vision as an example. The creation and assimilation of visual capabilities that surpass human vision in cameras, have been relatively easy. However, the more complex part was to add AI technology to the cameras. One such project is Landing.ai (https://www.landing.ai/) [6], formed by Andrew Ng, a globally recognized leader in AI. This startup focuses on solving manufacturing problems such as Quality Control (QC). It offers machine-vision tools to detect microscopic defects that may be found in products such as circuit boards, and which the human eye simply cannot detect.

Another recent and interesting project deals with machine touch. In the paper "Learning Dexterous In-Hand Manipulation" [7], a team of researchers and engineers at OpenAI have demonstrated that in-hand manipulation skills learned with reinforcement learning in a simulator can evolve into a fairly high level of dexterity of the robotic hand. As they explain: "This is possible due to extensive randomizations of the simulator, large-scale distributed training infrastructure, policies with memory, and a choice of sensing modalities which can be modelled in the simulator." (p. 15). The researchers' method "did not rely on any human demonstrations, but many behaviors found in human manipulation emerge naturally, including finger gaiting, multi-finger coordination, and the controlled use of gravity". (p. 1).

Cognitive capabilities—"A range of capabilities is included in this category including recognizing known patterns and categories (other than through sensory perception); creating and recognizing novel patterns and categories; logical reasoning and problem solving using contextual information and increasingly complex input variables; optimization and planning to achieve specific objectives given various constraints; creating diverse and novel ideas or a novel combination of ideas; information retrieval, which involves searching and retrieving information from a large range of sources; coordination with multiple agents, which involves interacting with other machines and with humans to coordinate group activity; and output articulation and presentation, which involves delivering outputs other than through natural language. These could be automated production of pictures, diagrams, graphs, or mixed media presentations." [5] (p. 34).

By using these capabilities, AI can amplify our own abilities: "Artificial intelligence can boost our analytic and decision-making abilities by providing the right information at the right time. But it can also heighten creativity". [8] (Para. 13). Consider for example Autodesk's Dreamcatcher AI which enhances the imagination of designers. As explained in the company's website:

"Dreamcatcher is a generative design system that enables designers to craft a definition of their design problem through goals and constraints. This information is used to synthesize alternative solutions that meet the objectives. Designers are able to explore trade-offs between many alternative approaches and select design solutions for manufacture." (https://autodeskresearch.com/projects/dreamcatcher) [9].

Some of the cognitive capabilities have achieved human level performance "such as recognizing simple/complex known patterns and categories other than sensory perception; Search and retrieve information from a large scale of sources—breadth, depth, and degree of integration." [5] (p. 35). However, other capabilities are currently below median performance such as create and recognize new patterns/categories; solve problems in an organized way using contextual information and increasingly complex input variables other than optimization and planning; create diverse and novel ideas, or novel combinations of ideas [5].

Natural language processing—"This consists of two distinct parts: natural language generation, which is the ability to deliver spoken messages, including with nuanced human interaction and gestures, and natural language understanding, which is the comprehension of language and nuanced linguistic communication in all its rich complexity." [5] (p. 34). As for Natural language generation, although there is progress in this area (such as Google duplex), the levels of performance according to the report are at best median. When it comes to Natural Language understanding, there is a long way ahead of us.

Yet, an example for an effective implementation of these capabilities (and more) is Aida (http://aidatech.io/) [10], a virtual assistant that is being used by SEB, a major Swedish bank. Aida interacts with masses of customers through natural-language conversations, and therefore has access to vast amounts of data. "This way she can answer many frequently asked questions, such as how to open an account or make cross-border payments. She can also ask callers follow-up questions to solve their problems, and she's able to analyze a caller's tone of voice and use that information to provide better service later." [8] (para. 16).

Physical capabilities—This includes gross motor skills, navigation (these two have reached human level performance), fine motor skills and mobility (these are more difficult and hence the performance levels are currently still median and below). "These capabilities could be implemented by robots or other machines manipulating objects with dexterity and sensitivity, moving objects with multidimensional motor skills, autonomously navigating in various environments and moving within and across various environments and terrain." [5] (p. 35).

While AIs like Cortana are essentially digital entities, there are other applications where "intelligence is embodied in a robot that **augments** a human worker. With their sophisticated sensors, motors, and actuators, AI-enabled machines can now recognize people and objects and work safely alongside humans in factories, warehouses, and laboratories." [8] (para. 16, our emphasis).

"Cobots" are probably the best example here. Collaborative robots, as Gonzalez [11] explains, "excel because they can function in areas of work previously occupied only by their human counterparts. They are designed with inherent safety features like force feedback and collision detection, making them safe to work right next to human operators." (para. 2).

Based on a white paper that Universal Robots—one of the leading companies in the robot market—has published [12], Gonzalez lists the seven most common applications for Cobots. One of them, for example, is "pick and place": "A pick and place task is any in which a workpiece is picked up and placed in a different location. This could mean a packaging function or a sort function from a tray or conveyor; the later [sic] often requires advanced vision systems." [11] (para. 3).

Social and emotional capabilities—"This consists of three types of capability: social and emotional **sensing**, which involves identifying a person's social and emotional state; social and emotional **reasoning**, which entails accurately drawing conclusions based on a person's social and emotional state, and determining an appropriate response; and social and emotional **action**, which is the production of an appropriate social or emotional response, both in words and through body language." [5] (p. 34).

Let us Consider Mattersight as an example. The company provides a highly sophisticated data analysis system that tracks customer' responses on the telephone. The software analyzes varied communicative micro-features such as, tone, volume, word choice, pauses, and so on. Then, in a matter of a few seconds, AI algorithms interpret these features, compare them to the company's databases, and come up with a personality profile for each customer. Based on this profile, the customer will be referred to the most appropriate service agent for him [13].

To sum-up the report, Manyika et al. [5] notes that from a mechanical point of view, they are fairly certain that perfection can be achieved. Because, already today, through deep reinforcement learning for example, robots can untie shoelaces and remove a nail from the back of a hammer. However, from the cognitive point of view, although the robot's "intelligence", has progressed, this is still where the greatest technical challenge lie:

"While machines can be trained to perform a range of cognitive tasks, they remain limited. They are not yet good at putting knowledge into context, let alone improvising. They have little of the common sense that is the essence of human experience and emotion. They struggle to operate without a pre-defined methodology. They are far more literal than people, and poor at picking up social or emotional cues. They generally cannot detect whether a customer is upset at a hospital bill or a death in the family, and for now, they cannot answer "What do you think about the people in this photograph?" or other open-ended questions. They can tell jokes without really understanding them. They don't feel humiliation, fear, pride, anger, or happiness. They also struggle with disambiguation, unsure whether a mention of the word "mercury" refers to a planet, a metal, or the winged god of Roman mythology. Moreover, while machines can replicate individual performance capabilities such as fine motor skills or navigation, much work remains to be done integrating these different capabilities into holistic solutions where everything works together seamlessly." (pp. 26–27).

3. AI Goes Hand in Hand with Our Understanding of Ourselves

Singularity is based on several assumptions: first, that there is a clear notion of what is human intelligence; and second, that AI can decrease the gap between human intelligence and machine intelligence. However, both of these assumptions are not clear yet. What is becoming more and more apparent is that AI goes hand in hand with our understanding of our own human intelligence and behavior.

"Intelligence" is a complex and multifaceted phenomenon that has for years interested researchers from countless fields of study. Among others, intelligence is studied from psychological, biological, economical, statistical, engineering, and neurological perspectives. New insights emerge over time from the various disciplines, many of which are adopted into the science of AI and contribute to its development and progress. The most striking example is the special and fruitful interrelationship between artificial intelligence and cognitive science.

Cognitive science and artificial intelligence arose at about the same time, in the late 1950s, and grew out of two main developments: "(1) the invention of computers and the attempts soon thereafter to design programs that could do the kinds of tasks that humans do, and (2) the development of information-processing psychology, later called cognitive psychology, which attempted to specify the internal processing involved in perception, memory, and thought. Cognitive science was a synthesis of the two, concerned both with the details of human cognitive processing and with the computational modeling of those processes." [14] (p. 1).

What AI does best is **analyze, categorize, and find the relationships** between large amounts of data, quickly and very effectively, coming up with highly accurate predictions. These capabilities, as Collins and Smith explained [14], were the outcome of three foci that have turned out to be three major bases for progress in AI: **formalisms**—such as mean-ends analysis, which are standard methods for representing and implementing cognitive processes; **tools or languages** for building intelligent programs such as John McCarthy's LISP [15]; and **programs**—beginning with the Dendral project [16], the first expert system to allow the formation of a scientific hypothesis.

"By contrast, psychology historically has made progress mainly by accumulating empirical phenomena and data, with far less emphasis on theorizing of the sort found in artificial intelligence. More specifically, psychological theories have tended to be constructed just to explain data in some experimental paradigm, and have tended to be lacking a well-founded mechanism, probably a relic of the behavioristic or stimulus-response approach that dominated psychology from the 1920s through the 1950s." [14] (p. 2).

At first, AI was mistakenly identified with the mechanical psychological viewpoint of behaviorism. The physicalism of stimulus and response looked similar to the action of computers, a reduction of man into its gears [17]. In psychology, a more 'humanistic' view of the science was demanded. It was agreed by all 'humanistic' psychologists that a good theory should be irreducible, its terms cannot be reduced to simple physical constituents, and that terms such as 'intention' should have a major part in the theory. Moreover, any action should have some meaning to the actor, and the meaning should be subjective. The 'humanistic' approach to psychology was a scientific revolution against positivistic psychology (in the Kuhnian sense) [17] (p. 396). It turned out that AI came to be very similar to the 'humanistic' viewpoint. Both AI and cognitive science were beginning to ask similar questions and to use many similar terms.

What was needed for a science of cognition was a much richer notion of knowledge representation and process mechanisms; and that is what artificial intelligence has provided. Cognitive psychologists gained a rich set of formalisms to use in characterizing human cognition [14] (p. 2). Some of the early and most important formalisms were means-ends analysis [18], Discrimination nets [19], Semantic networks [20], Frames and scripts [21,22], Production systems [23], Semantic primitives [24,25], Incremental qualitative analysis [26]. Through the years a wide range of formalisms were developed for analyzing human cognition, and many of them are still in use today.

Moreover, artificial intelligence has become a kind of theoretical psychology. Researchers who sought to develop a psychological theory could become artificial intelligence researchers without making their marks as experimentalists. Thus, as in physics, two branches of psychology were formed—experimental and theoretical—and cognitive science has become the interface where theorists and experimentalists sort things out [14].

Boden [17] suggests that AI can be used as a test-lab for cognitive science. It raises and exposes psychological questions that were deeply implicit. It suggests new terms, ideas and questions that were otherwise hidden. In that sense we dare say that computation is playing the role of language for cognitive science. Similar to the role of mathematics in physics, computation has become a language for constructing theories of the mind. Computation is a formal language that imposes a set of constraints on the kind of theories that can be constructed. But unlike mathematics, it has several advantages for constructing psychological theories: while mathematical models are often static, computational models are inherently process-oriented; while mathematical models, particularly in psychology, are content-independent, computational models can be content-dependent; and while computational models are inherently goal-oriented, mathematics is not. [14].

Questions that we should ask includes; is the use of the same terms and the same language in AI and cognitive sciences only an analogy? Could it imply something deeper? Can we insert true 'intention' and true 'meaning' into computer agents? How can we define such terms in AI? In fact, this is the main question of strong AI. This would bring AI and cognitive science much closer.

In an attempt to answer these questions, we refer to the viewpoint of Dennett [27]. Let's define the notion of 'meaning'; to put things very simplistically, we will say that an action of a computer agent has a 'meaning' (for the agent) if the action is changes some part of its environment and the agent can sense that change. For example, if the agent is a ribosome, then the transcription of an RNA into a series of amino-acids, later to become a protein, has a meaning since the protein has some function in changing the agent's environment. The action of the ribosome has a 'meaning' in the cytoplasm environment. Similarly, we can embed a 'meaning' in computer agents. It was suggested by Dennett that we human can insert a derived 'intention' in computers, and computers can derive a lower type of 'intention' in other computers. This was also brought up years ago by Minsky [28], using a different language.

It was suggested by Boden [17] that we can bridge the gap between 'humanistic' approach to cognitive science (in the sense discussed above) and physical mechanism. The way to do so is by introducing an inner representation of the self into the computer. Intentionality and meaning could be aimed (given a context) into this inner representation; the reduction or mechanism of the intentionality will be enabled by the design or architecture of the inner representation. Hence, in order to describe what is going on in the computer, the language of intentionality will be the most appropriate, in the same sense that we talk about our dog's intentions when we wish to describe or explain its behavior, without the need for a behavioristic language, or other physical terms. It will not be 'natural' or efficient to describe the action of the computer in the language of the state of its switches, we will say that this particular action was 'intended for' to comply with the 'state of mind' that the computer had. This sounds a somewhat pretentious goal, however it is based on the assumption that any future advancement in AI must stand on a basic cognitive architecture, much more basic and deeper than what we have today.

Most of the recent progress in AI have been driven by deep neural networks and these are related to the "connectionist" view of human intelligence. Connectionist theories essentially perceive learning—human and artificial—as rooted in interconnected networks of simple units, either real neurons or artificial ones, which detect patterns in large amounts of data. Thus, some in the machine learning field are looking to psychological research on human learning and cognition to help take AI to that next level. Although the concept of neural networks has existed since the 1940s, only today, due to an enormous increase in computing power and the amount and type of data available to analyze, deep neural networks have become increasingly powerful, useful and ubiquitous [29].

The theory of consciousness was recently investigated by AI researchers. It was suggested by Dennett [27] that consciousness is an emergent property of many small processes, or agents, each struggle for its homogeneity. Consciousness is not a stage with spotlights in which all subconscious processes are the audience. Consciousness is a dynamical arena where many agents appear and soon disappear. It resembles an evolutionary process occurring in a very short timescale [30]. On this very basis a few AI models were suggested, the Copycat model [31] and its more advanced Learning Intelligent Distribution Agent (LIDA) [32] model. These two are examples of a strong reciprocal interaction between AI and cognitive science.

Similar reciprocal relationships are now beginning to form between social sciences and artificial intelligence to become the field of artificial social intelligence (ASI). ASI is an interdisciplinary science, which was introduced years ago by Brent and others [33], and is only now becoming prevalent. ASI is a new challenge for social science and a new arena for the science of AI. It deals with the formalization of delicate social interactions, using it in AI to implement social behavior into robots. The prospects for social scientists were suggested years ago by Anderson [34]:

"It is time for sociology to break its intellectual isolation and participate in the cognitivist rethinking of human action, and to avail itself of theoretical ideas, techniques and tools that have been developed in AI and cognitive science" (p. 20).

"My argument is that sociologists have a great deal to learn from these disciplines, and that the adoption of concepts, methods and tools from them would change sociologists working habits [. . .]" (p. 215).

4. ASI, A New Challenge

While artificial cognitive intelligence has become a well-established and significant field of research, and has been heavily invested by both cognitive and artificial intelligence researchers, artificial social intelligence is in its early stages and has great potential for the advancement of smart machines in a new and essential way.

While cognitive artificial intelligence scientists "essentially view the mind as something associated with a single organism, a single computational system, social psychologists have long recognized that this is just an approximation. In reality the mind is social, it exists, not in isolated individuals, but in individuals embedded in social and cultural systems." [35] (p. 24).

It is well established now that there are sets of brain regions that are dedicated to social cognition. It was first shown on primates [36] and later on humans [37]. As Frith [38] explains: "The function of the social brain is to enable us to make predictions during social interactions." (p. 67). The social brain includes a variety of mechanisms, such as the amygdala which is activated in case of fear. It is also connected with the mechanism of prejudice, stereotyping, associating values with stimuli. It concerns both people—individual and group—and objects. Another such mechanism is the medial prefrontal cortex, which is connected with the understanding of the other's behavior in terms of its mental state, with long term dispositions and attitudes, and with self-perception about long term attitudes.

From the social point of view, Mead [39], in his book, *Mind, Self and Society*, defines the "social organism" as "a social group of individual organisms" (p. 130), or in modern language, as an emergent phenomenon. This means that each individual, as an organism in itself, is also a part of a larger system, the social organism. Hence, each individual's act must be understood within the context of some social act that involve other individuals. The social act is therefore viewed as a dynamic and complex system within which the individual is situated. As such, the social 'organism' actually defines the individual acts, that is, within it these acts become meaningful.

In his book *Artificial Experts* [40], Collins argues similarly **that intelligence cannot be defined without considering social interactions**. This is because "[. . .] the locus of knowledge appears to be not the individual but the social group; what we are as individuals is but a symptom of the groups in which the irreducible quantum of knowledge is located. Contrary to the usual reductionist model of the social sciences, it is the individual who is made of social groups." (p. 6).

Our intelligence, as Yudkowsky [41] clarifies, "includes the ability to model social realities consisting of other humans, and the ability to predict and manipulate the internal reality of the mind." (p. 389). Another way to put it is through Mead's concept of the 'Generalized other' [39]. As Dodds, Lawrence & Valsiner [42] explain, "**to take the role of the other involves the importation of the social into the personal**, and this activity is crucial for the development of self-consciousness and the ability to operate in the social world. It describes how perspectives, attitudes and roles of a group are incorporated into the individual's own thinking in a way that is distinct from the transmission of social rules, and in a way that can account for the possibility of change in both person and society." (p. 495, our emphasis).

Hence, as Collins [40] argues, "The organism into which the intelligent computer supposed to fit is not a human being but a much larger organism; a social group. The intelligent computer is meant to counterfeit the performance of a whole human being within a social group, not a human being's brain. **An artificial intelligence is a '*social prosthesis*'.**" (p. 14, our emphasis).

All the above suggests the emergence of a new interdisciplinary discipline. The main concern of this new field of science is the formalization of delicate social modules, using them in AI to implement social awareness (perhaps a type of social common sense understanding) and social behavior into robots. Because of the dynamic nature of social interactions, these ASI systems face difficult challenges, some of which are not even predictable. In order to address these challenges, ASI systems will have to be dynamic by continuously reviewing and evolving their interaction strategies in order to adapt to new social situations. Moreover, it is essential to examine and assess these strategies in as many contexts as possible, in which ongoing, continuous interactions are taking place.

For making ASI come true, there are some fundamental steps which needs to be solved [43]. Firstly, there is a need to discover the principles of socio-culture interactions in which the ASI system could have a role. In order to formulate those principles there is considerable importance for conducting large data-driven studies aimed at validating these principles, as well as identifying and characterizing new behavioral traits. Such studies are already being conducted, using the enormous amounts of socially grounded user data generated and highly available from social media; as well as the significant advancements in machine learning and the wide variety of data-analysis techniques. One such project is "Mark my words!" [44]. This project demonstrates the psycholinguistic theory of communication accommodation according to which participants in conversations tend to adapt to the communicative behavior patterns of those with whom they converse. The researches have shown "that the hypothesis of linguistic style accommodation can be confirmed in a real life, large scale dataset of Twitter conversations." (p. 754). A probabilistic framework was developed, which allowed the researchers to measure "accommodation and, importantly, to distinguish effects of style accommodation from those of homophily and topic-accommodation." [44].

Once the relevant socio-cultural principles have been extracted and defined, the next step will be to understand how they can be assimilated into ASI systems such as chatbots, recommender systems, autonomous cars, etc. One such system is the *virtual receptionist*, "which keeps track of users attention and engagement through visual cues (such as gaze tracking, head orientation etc.) to initiate the interaction at the most appropriate moment [45]. Further, it can also make use of hesitation (e.g., "hmmm ... uhhh") to attract the attention of the user, buy time for processing or even to indicate uncertainty in the response [46]." [43] (para. 6).

ASI systems have no clear definition of goals, there is no specific task the machine is oriented towards. In a sense, the machine's social behavior is the goal. In other words, it is impossible to defined clear goals in advance, and these may even emerge dynamically. This means that measurement and evaluation methods are very difficult to apply to a socio-cultural intelligence of such a system. This is one of the biggest challenges the ASI field has to deal with.

5. AGI, An Overview, Is It Enough?

An important concept to dwell on is that of artificial general intelligence (AGI). AGI constitute a new step towards strong AI. General intelligence is not a fully well-defined term, but it has a qualitative meaning: “What is meant by AGI is, loosely speaking, AI systems that possess a reasonable degree of self-understanding and autonomous self-control, and have the ability to solve a variety of complex problems in a variety of contexts, and to learn to solve new problems that they didn’t know about at the time of their creation.” [35] (p. VI).

There is a clear distinction between AGI and narrow AI research. The latter is aimed at creating programs that specialize in performing specific tasks, such as ordering online shopping, playing GO, diagnosing diseases or driving a car. But, despite their great importance and popularity, narrow AIs core problem is that “they are inherently narrow (narrow by design) and fixed. Whatever capabilities they have, are pretty much frozen in time. It is true that narrow AI can be designed to allow for some limited learning or adaptation once deployed, but this is actually quite rare. Typically, in order to change or expand functionality requires either additional programming, or retraining (and testing) with a new dataset.” [47] (para. 4–5).

Intelligence, in general, “implies an ability to acquire and apply knowledge, and to reason and think, in a variety of domains” [48] (p. 15). In other words, intelligence in its essence has a large and dynamical spectrum.

“Narrow AI systems cannot adapt dynamically to novel situations—be it new perceptual cues or situations; or new words, phrases, products, business rules, goals, responses, requirements, etc. However, in the real world things change all the time, and intelligence is by definition the ability to effectively deal with change.” [47] (para. 6).

Artificial general intelligence requires the above characteristics. It must be capable of performing various tasks in different contexts, making generalizations and tapping from existing knowledge in a given context to another. Hence, as Voss [47] explains, “it must embody at least the following essential abilities:

(1) To autonomously and interactively acquire new knowledge and skills, in real time. This includes one-shot learning—i.e., learning something new from a single example.
(2) To truly understand language, have meaningful conversation, and be able to reason contextually, logically and abstractly. Moreover, it must be able to explain its conclusions.
(3) To remember recent events and interactions (short-term memory), and to understand the context and purpose of actions, including those of other actors (theory of mind).
(4) To proactively use existing knowledge and skills to accelerate learning (transfer learning).
(5) To generalize existing knowledge by forming abstractions and ontologies (knowledge hierarchies).
(6) To dynamically manage multiple, potentially conflicting goals and priorities, and to select the appropriate input stimuli and to focus on relevant tasks (focus and selection).
(7) To recognize and appropriately respond to human emotions (have EQ, emotional intelligence), as well as to take its own cognitive states—such as surprise, uncertainty or confusion—into account (introspection).
(8) Crucially, to be able to do all of the above with limited knowledge, computational power, and time. For example, when confronted with a new situation in the real world, one cannot afford to wait to re-train a massive neural network over several days on a specialized supercomputer. (para. 12).

In conclusion, general intelligence is a complex phenomenon that emerges from the integration of several essential components. “On the structural side, the system must integrate sense inputs, memory, and actuators, while on the functional side various learning, recognition, recall and action capabilities must operate seamlessly on a wide range of static and dynamic patterns. In addition, these cognitive abilities must be conceptual and contextual—they must be able to generalize knowledge, and interpret it against different backgrounds.” [49] (p. 147).

From the point of view of strategy and methodology AGI sometimes uses a top down approach on cognition, as Wang and Goertzel [50] explains, "An AGI project often starts with a blueprint of a whole system, attempting to capture intelligence as a whole. Such a blueprint is often called an "architecture"." (p. 5).

Cognitive architecture (CA) research "models the main factors participated in our thinking and decision and concentrates on the relationships among them. In computer science, CA mostly refers to the computational model simulating human's cognitive and behavioral characteristics. Despite a category of loose definition, CAs usually deal with relatively large software systems that have numerous heterogeneous parts and subcomponents. Typically, many of these architectures are built to control artificial agents, which run both in virtual worlds and physical robots." [51] (p. 1).

Symbolic systems are one important type of cognitive architecture. "This type of agents maintains a consistent knowledge base by representing the environment as symbols." [51] (p. 2). Some of the most ambitious AGI-oriented projects in the history of the field were in the symbolic-AI paradigm. One such famous project is the General Problem Solver [52], which used heuristic search (means-ends analysis) to solve problems. Another famous effort was the CYC project [53]. The project's aim was to create human-like AI by collecting and encoding all human common sense knowledge in first order logic. Alan Newell's SOAR project [54,55] was an attempt to create unified cognition theories, based on "logic-style knowledge representation, mental activity as problem-solving carried out by an assemblage of heuristics, etc." [35] (p. 3). However, the system was not constructed to be fully autonomous or to have self-understanding [35].

These and other early attempts failed to reach their original goals, and in the view of most AI researchers, failed to make dramatic conceptual or practical progress toward their goals. Some (GPS for example) failed because of exponential growth in computational complexity. However, more contemporary AGI studies and projects offer new approaches, combining the previous knowledge—both theories and research methods—accumulated in the field.

One such integrative scheme described by Pennachin and Goertzel [35], was given the name 'Novamente'. This scheme involves taking elements from various approaches and creating an integrated and interactive system. However, as the two explain: "This makes sense if you believe that the different AI approaches each capture some aspect of the mind uniquely well. But the integration can be done in many different ways. It is not workable to simply create a modular system with modules embodying different AI paradigms: the different approaches are too different in too many ways. Instead one must create a unified knowledge representation and dynamics framework, and figure out how to manifest the core ideas of the various AI paradigms within the universal framework." (p. 5).

In their paper, "Novamente: an integrative architecture for Artificial Intelligence" [56], Goertzel et al. suggest such an integrative AI software system. The Novamente design incorporates evolutionary programming, symbolic logic, agent systems, and probabilistic reasoning. The authors clarify that "in principle, integrative AI could be conducted in two ways: Loose integration, in which different narrow AI techniques reside in separate software processes or software modules, and exchange the results of their analysis with each other. Tight integration, in which multiple narrow AI processes interact in real-time on the same evolving integrative data store, and dynamically affect one another's parameters and control schemata. Novamente is based on a distributed software architecture, in which a distributed processing framework called DINI (Distributed Integrative Intelligence) is used to bind together databases, information-gathering processes, user interfaces, and "analytical clusters" consisting of tightly-integrated AI processes." (p. 2).

Novamente is extremely innovative in its overall architecture, which seeks to deal with the difficulty of creating a "whole brain" in a completely new and direct way. The basic principles on which the design of the system is founded are derived from the "psynet model"—an innovative complex-systems theory of mind—which was developed by Goertzel [57–61]. "What the psynet model has led us to is not a conventional AI program, nor a conventional multi-agent-system framework. Rather, we are talking about an autonomous, self-organizing, self-evolving AGI system, with its own understanding of the world, and the ability to relate to humans on a "mind-to-mind" rather than a "software-program-to-mind" level." [35] (pp. 64–65).

Another interesting project is the Learning Intelligent Distribution Agent (LIDA) [62]. The LIDA architecture is presented as a working model of cognition, a Cognitive Architecture, which was designed to be consistent with what is known from cognitive sciences and neuroscience. Ramamurthy et al. argue "that such working models are broad in scope and could address real world problems in comparison to experimentally based models which focus on specific pieces of cognition. [. . .] A LIDA based cognitive robot or software agent will be capable of multiple learning mechanisms. With artificial feelings and emotions as primary motivators and learning facilitators, such systems will 'live' through a developmental period during which they will learn in multiple ways to act in an effective, human-like manner in complex, dynamic, and unpredictable environments." (p. 1).

In a nutshell, LIDA is a modified version of the old COPYCAT architecture suggested years ago by Hofstadter [31]. It is based on the attempt to understand consciousness as a working space for many agents. The agents compete one another and those that dominate the workspace are identified as the ones that constitute our awareness. The process is dynamic, information flows in from the environment, and action is decided by a set of heuristics, which are themselves dynamic.

The LIDA architecture is partly symbolic and partly connectionist; part of the architecture "is composed of entities at a relatively high level of abstraction, such as behaviors, message-type nodes, emotions, etc., and partly of low-level codelets (small pieces of code). LIDA's primary mechanisms are perception, episodic memory, procedural memory, and action selection." [62] (p. 1).

With the design of three continually active incremental learning mechanisms—perceptual learning, episodic learning and procedural learning—the researchers have laid the foundation for a working model of cognition that produces a cognitive architecture capable of human like learning. As the authors [62] explain:

"The architecture can be applied to control autonomous software agents as well as autonomous robots "living" and acting in a reasonably complex environment. The perceptual learning mechanism allows each agent controlled by the LIDA architecture to be suitably equipped so as to construct its own ontology and representation of its world, be it artificial or real. And then, an agent controlled by the LIDA architecture can also learn from its experiences, via the episodic learning mechanism. Finally, with procedural learning, the agent is capable of learning new ways to accomplish new tasks by creating new actions and action sequences. With feelings and emotions serving as primary motivators and learning facilitators, every action, exogenous and endogenous taken by an agent controlled with the LIDA architecture is self-motivated." (p. 6).

A third project worth mentioning is Schmidhuber's Gödel Machines [63]. Schmidhuber describe these machines as "the first class of mathematically rigorous, general, fully self-referential, self-improving, optimally efficient problem solvers. Inspired by Kurt Gödel's celebrated self-referential formulas (1931), such a problem solver rewrites any part of its own code as soon as it has found a proof that the rewrite is *useful,* where the problem-dependent utility function and the hardware and the entire initial code are described by axioms encoded in an initial proof searcher which is also part of the initial code. The searcher systematically and in an asymptotically optimally efficient way tests computable *proof techniques* (programs whose outputs are proofs) until it finds a provably useful, computable self-rewrite." (p. 1).

In other words, the Gödel machines "are universal problem solving systems that interact with some (partially observable) environment and can in principle modify themselves without essential limits apart from the limits of computability. Their initial algorithm is not hardwired; it can completely rewrite itself, but only if a proof searcher embedded within the initial algorithm can first prove that the rewrite is useful, given a formalized utility function reflecting computation time and expected future success (e.g., rewards)." (p. 2).

A completely different approach to AGI suggests imitating the complex architecture of the human brain and creating its exact digital simulation. However, this method is questionable since the brain has not been fully deciphered yet. Another, more abstract way to create AGI is to follow cognitive psychology research and to emulate the human mind. A third way is to create AGI by emulating properties of both aspects—brain and mind. But, as Wang [64] stresses, the main issue is not "whether to learn from the human brain/mind (the answer is always "yes", since it is the best-known form of intelligence), or whether to idealize and simplify the knowledge obtained from the human brain/mind (the answer is also always "yes", since a computer cannot become identical to the brain in all aspects), but on *where* to focus and *how much* to abstract and generalize." (pp. 212–213).

One of the unsolved problems of AGI research is our lack of understanding of the definition of "Generalization", but what Perez [65] suggests "is that our measure of intelligence be tied to our measure of social interaction." (para. 7). Perez calls his new definition for generalization "Conversational Cognition" and as he explains:

"An ecological approach to cognition is based on an autonomous system that learns by interacting with its environment. Generalization in this regard is related to how effectively automation is able to **anticipate** contextual changes in an environment and perform the required context switches to ensure high predictability. The focus is not just in recognizing chunks of ideas, but also being able to recognize the relationship of these chunks with other chunks. There is an added emphasis on recognizing and predicting the opportunities of change in context." (para. 11).

The most sophisticated form of generalization that exists demands the need to perform conversations. Moreover, Perez [65] clarifies that this conversation "is not confined only to an inanimate environment with deterministic behavior. [. . .] we need to explore conversation for computation, autonomy and social dimensions. [. . .] The social environment will likely be the most sophisticated system in that it may demand understanding the nuisances of human behavior. This may include complex behavior such as deception, sarcasm and negotiation." (para. 13, 14).

Another critical aspect of social survival is the requirement for cooperative behavior. But as Perez [65] argues, effective prediction of an environment is an insufficient skill to achieve cooperative behavior. The development of language is a fundamental skill, and conversations are the highest reflection of intelligence. "They require the cognitive capabilities of memory, conflict detection and resolution, analogy, generalization and innovation." (para. 15). But at the same time it is important to keep in mind that languages are not static—they evolve over time with new concepts.

Moreover, Perez [65] clarifies that "effective conversation requires not only understanding an external party but also the communication of an automaton's inner model. In other words, this conversation requires the appropriate contextualized communication that anticipates the cognitive capabilities of other conversing entities. Good conversation requires good listening skills as well as the ability to assess the current knowledge of a participant and performing the necessary adjustment to convey information that a participant can relate to." (para. 16). For Perez, the ability to effectively perform a conversation with the environment is the essence of AGI. Interestingly enough, what most AGI research avoids is the reality that an environment is intrinsically social—i.e., that there exist other intelligences.

As we have argued above, we believe that the next step to take to make human and machine intelligence come closer together, is to focus on the social aspect of human intelligence and on the ways to integrate social behavior in machines.

6. Embodiment

One of the biggest challenges for AI is the challenge of embodied cognition. If AI could surpass this hurdle it will be very close to true human intelligence. Embodied cognition was first presented as the main reason why AI is impossible. We propose to view embodied cognition as a step towards better AI and not as an imposition. Let us make a small detour to the history and philosophy of computation.

Dreyfus [66] claimed that true AI is impossible since it implicitly assumes that human intelligence is symbolic in its essence. Some AI researchers are attempting to build a context-free machine that manipulates symbols, assuming the human mind works similarly. Dreyfus claimed that the symbolic conjecture is a fault, basing his arguments primarily on philosophical grounds. AI assumes that we have a type of 'knowledge representation' in our brain, a representation of the world, this idea is based on Descartes' theory, and has a long tradition. Moreover, Descartes claimed that there is a duality, a separation between our body and our mind, therefore the mind cannot be embodied. So far, claimed Dreyfus [66], all AI research is based on these assumptions that we have a model of the world in our mind and that the mind is separated from the body.

Could these assumptions be wrong? Could it be that some part of our intelligence in embedded in our body? We interact with the world with our body, we perceive with our body, we sense with our body. Could it be that symbolic intelligence is not enough?

For Dreyfus [66], embodiment is enrooted in the deep philosophical grounds of existentialism. Existentialism discusses the notions of involvement and detachment. Most of the time, humans are involved in the world, they interact, they solve practical problems, they are involved in everyday coping, finding their way about in the world. However, when things become difficult, the individual retracts into detachment. For most of the things you do, there is no need for any type of awareness; while climbing stairs you do not think about the next step. If you do you will probably fall. Only when the stairs are too steep, you might consider your next step, and then you retract to a state of detachment. For Heidegger [67] there is the World where we live, where everything has 'meaning', and there is the Universe where detachment and science lives. Science is the outcome of our involvement in the world, and not the other way around. Science cannot explain the 'meaning' of things. Existentialism is therefore the opposite of Descartes' dualism.

Part of our intelligence is therefore somewhere in our bodily interactions with the world. In addition to our senses of sight, smell, hearing etc., we have 'senses' of time, space, surroundings, etc. The discovery of neural place cells [68–70] emphasizes the embodiment of our sense of space. A good example to illustrate embodiment is the proven connection between movement and intelligence in baby development. Free movement, such as rolling, crawling, sitting, walking, jumping, etc., is associated with the development of the frontal cortex, the area where higher-order thinking occurs [71].

We can coin the above set of 'senses' as 'environmental intelligence'. The question is how much of our intelligence is grounded in our body, and how much is 'context free' in our mind? If we had no body, could we think the same way we think? Could we think at all? Would we have anything to think about? What is the connection between our 'environmental intelligence' and our 'context free symbolic manipulation intelligence'?

Dreyfus [66] thought that neural network computations are indeed a step in the right direction. It is an attempt to formalize our perceptions in terms of virtual neurons which have some parallel in our brain and body.

There were computer scientists that took Dreyfus' stand seriously. Brooks [72] came up with the idea that there is no need for knowledge representation at all: "The key observation is that the world is its own best model" (p. 6).

Brooks' robots were able to wander around, to avoid obstacles and to perform several basic tasks. A more modern version would be an intelligent swarm, where a set of simple agents interact and can bring about some emergent property [73].

Brooks and Dennett cooperated in a project involving a humanoid named COG [74,75], where they tried to implement the above ideas by letting the COG computer (with a torso, camera eyes, one arm, and three fingers) interact with its environment, trying to learn new things as if it is a newborn. Brooks used a 'subsumption architecture' where several simple modules were competing for dominance. A few years later, the science of Embodied Cognition was born [76,77] and reached similar conclusions from a different point of view, the point of view of cognitive psychology.

The science of embodied cognition has several basic assumptions. First, humans have a set of atomic and primitive cognitive abilities that are embodied. These abilities build our perceptions of the world. Lakoff and Johnson [76] used the term 'image schema'. Such a schema is a small process, which therefore could also be dynamic. Furthermore, any higher feature, any concept that we form, is the result of aggregations of the above primitives. Finally, we think by using Metaphors and Frames and these are formed using associations of schemas.

As for Frames, many words have a large and natural context and cannot be understood without their context, for example prisoner, nurse, doctors, etc. These are the Frames. Frames were suggested in social science by Goffman [78], and they were also referred to in the context of AI by Minsky [21]. Minsky was also interested in issues such as: the symbolic aspect of vision, the relation of his theory of Frames to Piaget's theory of development, language understanding, scenarios etc. Dreyfus [66], on the other hand, stressed the fact that real frames are infinite in nature and could not be truly described in AI. This was coined the 'Frame Problem'.

Metaphors are formed using Hebbian learning [79]. In early childhood we learn to associate between several concepts, such as 'warm' and 'up', or 'cold' and 'down'. The more these associations of schemas are presented to us in childhood the stronger we relate the schemas. Later we use more elaborated metaphors to reason. We solve real situations by using homeomorphisms into a world of metaphors, solving the imaginary situation first. This is 'thinking by metaphors' [77].

To prove all of the above, embodied cognition scientists search for clues in language where they look for invariants. The existence of such invariants can imply that something deep, common to all languages, underlies. In many examples a word has several meanings, one is environmental and embodied, the other much more abstract. For example 'to grasp' is first of all 'to catch', however it also has the meaning of 'to understand'. We 'see' things in the sense of understanding, we talk about a 'warm' or 'cold' person, etc. Old proverbs are a good source for such examples.

Lakoff and Johnson's [77] argument is that the more abstract meaning is the secondary one, it is the derived meaning. We derive such meanings from the simpler ones, the embodied and primitive meanings. It is a bottom up process.

Artificial Social Intelligence is also concerned with the environment of the intelligent agent, in particular its social environment. However, there is a difference between the theory of ASI and Embodied Cognition. ASI is rooted in social science, and the notion of a 'social brain'. Embodied cognition is rooted in the intersection of cognitive sciences and linguistics, it is also based on philosophical grounds. Embodied cognition focuses on universal principles of understandings, or in other words, ***the*** cognitive architecture.

Can we follow the way from the embodied primitives upwards to the abstract concepts? Can we use AI to boost such research? This was attempted by several researchers; Regier [80] used the theory of deep neural networks (and recurrent neural networks) to identify the primitives of cognition used in relation to our sense of space. In deep neural networks we take very simple and primitive features and build upon the next layer of more complex features. This is a multi-step process. This bottom up process happens until a global pattern can be recognized and it resembles the process that was defined by Lakoff and Johnson [76] for schemas. Regier [80] was struggling with old methods of back propagation to tune his network. Today we can advance Regier's idea by using new techniques in Deep Neural Networks.

Another way in which we can implement embodiment cognition is by formalizing the idea of metaphors. To be able to use metaphors we need to enable the computer the capability to simulate a situation in which the machine itself resides. This was already done in the context of value alignment. Winfield, Blum and Liu [81] defined a 'consequence machine' that could simulate a situation, but could also observe itself in that simulation. The machine then had to decide on a 'moral' dilemma.

Embodied cognition is a new hurdle to overcome, it is the missing bridge between robotics and AI. It should not be thought of as an imposition on AI but as a new challenge.

7. The Way to Overcome Our Fears: Value Alignment

In an article called "How Do We Align Artificial Intelligence with Human Values?" [82], Conn explains that "highly autonomous AI systems should be designed so that their goals and behaviors can be assured to align with human values throughout their operation". (para. 5).

One of the main challenges in aligning AI with values is to understand (and to agree upon) what exactly these values are. There are many factors that must be taken into account which depend mainly on context—cultural, social, socioeconomic and more. It is also important to remember that humanity often does not agree on common values, and even when it does, social values tend to change over time.

Eliezer Yudkowsky offered the first attempt at explaining AI alignment in his seminal work on the topic, "Creating Friendly AI" [83], and he followed this up with a more nuanced description of alignment in "Coherent Extrapolated Volition" [84]. Nearly a decade later Stuart Russell began talking about the value alignment problem, giving AI alignment its name and motivating a broader interest in AI safety. Since then numerous researchers and organizations have worked on AI alignment to give a better understanding of the problem.

As Tegmark [85] explains: "aligning machine goals with our own involves three unsolved problems: making machines learn them, adopt them and retain them. AI can be created to have virtually any goal, but almost any sufficiently ambitious goal can lead to subgoals of self-preservation, resource acquisition and curiosity to understand the world better—the former two may potentially lead a superintelligence AI to cause problems for humans, and the latter may prevent it from retaining the goals we give it." (p. 389).

How to implement Value Alignment? Wilson and Daugherty [8] describe three critical roles that we, humans, need to perform:

Training: Developing 'personalities' for AI requires considerable training by diverse experts. For example, in order to create Cortana's personality, Microsoft's AI assistant, several human trainers such as a play writer, novelist and poet, spent hours in helping developers create a personality that is confident, helpful and not too 'bossy'. Apple's Siri is another example. Much time and effort was spent to create Siri with a hint of sassiness, as expected from an Apple product.

Creating AI with more complex and subtle human traits is sought after by new startups for AI assistants. Koko, a startup born out of the MIT Media Lab, has created an AI assistant that can display sympathy. For example, if a person is having a bad day, it will not just say 'I'm sorry to hear that', but will ask for more information and perhaps provide advice like 'tension could be harnessed into action and change' [86].

Explaining: As AI develops, it reaches results through processes that are unclear to users at times, a sort of internal 'black box'. Therefore, they require expert, industry specific 'explainers' for us to understand how AI reached a certain conclusion. This is especially critical in evidence-based industries such as medicine and law. A medical practitioner must receive an explanation of why an AI assistant gave a certain recommendation, what is the internal 'reasoning' that led to a decision. In a similar way, law enforcement investigating an autonomous vehicle accident, need experts to explain the AI's reasoning behind decisions that led to an accident [8].

Sustaining: AI also requires sustainers. Sustainers oversee and work on making sure AI is functioning as intended, in a safe and responsible manner. For example, a sustainer would make sure an autonomous car recognizes all human diversity and takes action not to risk or harm any human being. Other sustainers may be in charge of making sure AI is functioning within the desired ethical norms. For example, when analyzing big data to enhance user monetization, a sustainer would oversee that the process is using general statistical data and not specific and personal data (which may generate negative sentiment by users) to deduce its conclusions and actions [8].

The unique roles of human values presented here, have been linked to the workplace environment, but they are undoubtedly relevant to all spheres of life. As we claimed earlier, we are at the beginning of a new developmental stage in AI, the one of artificial social intelligence. Within this realm, new questions concerning human values may arise.

8. Summary

In his book *Technopoly* [87] Postman writes that "[. . .] once a technology is admitted, it plays out its hand; it does what it is designed to do. Our task is to understand what that design is—that is to say, when we admit a new technology to the culture, we must do so with our eyes wide open." (p. 7). For "technological change is neither additive nor subtractive. It is ecological. I mean 'ecological' in the same sense that the word is used by environmental scientists. One significant change generates total change." (p. 18).

AI is woven into our lives, changing our environment. The short history of AI has shown that developments in AI go hand in hand with our understanding of ourselves. Although there is still a long way to go before we can talk about a singularity point, it is almost clear that the next few steps in AI technology (e.g., ASI and AGI) will bring about a much more powerful machines, flexible enough to resemble human behavior.

A major part of human intelligence is social, we interact with others, we compete, we cooperate, we imitate, etc. A symbolic 'context free' intelligence cannot be complete without this social constituent. ASI is therefore necessary for building 'true' human intelligence.

One of the most complex challenges that AI faces is the issue of embodiment. At first, one should recognize that part of our intelligence is indeed embodied. We are physically situated in the world; we move in space, perceive, feel and communicate through our bodies. We argue that embodied cognition should be dealt with as a new challenge to AI and not as an imposition.

Funding: This reaserch did not receive any funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lévi-Strauss, C.; Weightman, J. *The Raw and the Cooked: Introduction to a Science of Mythology*; Harper & Row: New York, NY, USA, 1969; Volume 1.
2. Kang, M. *Sublime Dreams of Living Machines*; Harvard University Press: Cambridge, MA, USA, 2011.
3. Strate, L. *Media Ecology: An Approach to Understanding the Human Condition*; Peter Lang Publishing, Incorporated: Bern, Switzerland, 2017.
4. McLuhan, M. *Understanding Media: Extensions of Man*; McGraw Hill: New York, NY, USA, 1964.
5. Manyika, J.; Michael, C.; Mehai, M.; Jacques, B.; Katy, G.; Paul, W.; Martin, D. The Future That Works: Automation, Employment and Productivity. Mckinsey and Company Global Institute, January 2017. Available online: https://www.mckinsey.com/~{}/media/McKinsey/Featured%20Insights/Digital%20Disruption/Harnessing%20automation%20for%20a%20future%20that%20works/MGI-A-future-that-works_Full-report.ashx (accessed on 17 July 2018).
6. Landing.ai. Available online: https://www.landing.ai/ (accessed on 17 July 2018).
7. Andrychowicz, M.; Baker, B.; Chociej, M.; Jozefowicz, R.; McGrew, B.; Pachocki, J.; Petron, A.; Plappert, M.; Powell, G.; Ray, A.; et al. Learning Dexterous In-Hand Manipulation. OpenAI, 2018. Available online: https://arxiv.org/pdf/1808.00177.pdf (accessed on 1 October 2018).
8. Wilson, H.J.; Daugherty, P.R. Collaborative Intelligence: Humans and AI are Joining Forces. Harvard Business Review. July–August Issue. 2018. Available online: https://hbr.org/2018/07/collaborative-intelligence-humans-and-ai-are-joining-forces (accessed on 4 July 2018).
9. AutoDesk's Dreamcatcher. Available online: https://autodeskresearch.com/projects/dreamcatcher (accessed on 17 July 2018).
10. AIDA—Artificial Intelligence Driven Analytics. Available online: http://aidatech.io/ (accessed on 5 August 2018).
11. Gonzalez, C. Seven Common Applications for Cobots. MachineDesign. Available online: http://www.machinedesign.com/motion-control/7-common-applications-cobots (accessed on 18 July 2018).
12. Universal Robots. White Paper: An Introduction to Common Collaborative Robot Applications. 2018. Available online: https://cdn2.hubspot.net/hubfs/2631781/HQ%20Content%20and%20Enablers/HQ%20Enablers/White%20papers/Common%20Cobot%20Applications.pdf (accessed on 18 July 2018).
13. Stevens, G. AI Is over: This is Artificial Empathy. Kernel Magazine. 2013. Available online: https://kernelmag.dailydot.com/features/report/5910/ai-is-so-over-this-is-artificial-empathy/# (accessed on 17 October 2013).

14. Collins, A.; Smith, E.E. (Eds.) *Readings in Cognitive Science: A Perspective from Psychology and Artificial Intelligence*; Elsevier: Amsterdam, The Netherlands, 2013.
15. McCarthy, J. History of LISP. In *History of Programming Languages I*; ACM: New York, NY, USA, 1978; pp. 173–185.
16. Lindsay, R.K.; Buchanan, B.G.; Feigenbaum, E.A.; Lederberg, J. *Applications of Artificial Intelligence for Organic Chemistry: The DENDRAL Project*; McGraw-Hill Companies, Inc.: New York, NY, USA, 1980.
17. Boden, M.A. *Artificial Intelligence and the Natural Man*, 1st ed.; The MIT Press: London, UK, 1977.
18. Newell, A.; Simon, H.A. GPS a program that simulates human thoughts. In *Computer and Thoughts*; Feigenbaum, E.A., Feldman, J., Eds.; McGraw-Hill: New York, NY, USA, 1963.
19. Feigenbaum, E.A. The simulation of verbal learning behavior. In *Computers and Thought*; Feigenbaum, E.A., Feldman, L., Eds.; McGraw-Hill: New York, NY, USA, 1963; pp. 297–309.
20. Quillian, M.R. Semantic memory. In *Semantic Information Processing*; Minsky, M., Ed.; MIT Press: Cambridge, MA, USA, 1968; pp. 216–270.
21. Minsky, M. A framework for representing knowledge. In *The Psychology of Computer Vision*; Winston, P.H., Ed.; McGraw Hill: New York, NY, USA, 1975; pp. 211–277.
22. Schank, R.C.; Abelson, R.P. *Scripts, Plans, Goals, and Understanding*; Lawrence Erlbaum Associates: Hillsdale, NJ, USA, 1977.
23. Newell, A.; Simon, H.A. *Human Problem Solving*; Prentice-Hall: Englewood Cliffs, NJ, USA, 1972; Volume 104, No. 9.
24. Schank, R.C. Conceptual dependency: A theory of natural language understanding. *Cogn. Psychol.* **1972**, *3*, 552–631. [CrossRef]
25. Norman, D.A.; Rumelhart, D.E.; The LNR Research Group. *Explorations in Cognition*; W.H. Freeman: San Francisco, CA, USA, 1975.
26. De Kleer, J. The origin and resolution of ambiguities in causal arguments. In *Readings in Qualitative Reasoning about Physical Systems*; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 1990; pp. 624–630.
27. Dennett, D.C. From Bacteria to Bach and back: The Evolution of Minds. WW Norton & Company: New York, NY, USA, 2017.
28. Minsky, M. *The Emotion Machine*; Pantheon: New York, NY, USA, 2006; p. 56.
29. Winerman, L. Making a thinking machine. *Am. Psychol. Assoc.* **2018**, *49*, 4. Available online: https://www.apa.org/monitor/2018/04/cover-thinking-machine.aspx (accessed on 17 November 2018).
30. Calvin, W.H. *How Brains Think*; Basic Books: New York, NY, USA, 1996.
31. Hofstadter, D.R. *Fluid Concepts and Creative Analogies*; Basic Books: New York, NY, USA, 1995.
32. Faghihi, U.; Franklin, S. The LIDA model as a foundational architecture for AGI. In *Theoretical Foundations of Artificial General Intelligence*; Atlantis Press: Paris, France, 2012; pp. 103–121.
33. Bainbridge, W.S.; Brent, E.E.; Carley, K.M.; Heise, D.R.; Macy, M.W.; Markovsky, B.; Skvoretz, J. Artificial social intelligence. *Annu. Rev. Sociol.* **1994**, *20*, 407–436. [CrossRef]
34. Anderson, B. On artificial intelligence and theory construction in sociology. *J. Math. Sociol.* **1989**, *14*, 209–216. [CrossRef]
35. Pennachin, C.; Goertzel, B. Contemporary approaches to artificial general intelligence. In *Artificial General Intelligence*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 1–30.
36. Brothers, L. The social brain: A project for integrating primate behavior and neurophysiology in a new domain. In *Foundations in Social Neuroscience*; MIT Press: Cambridge, MA, USA, 2002; pp. 367–385.
37. Adolphs, R. Cognitive neuroscience of human social behaviour. *Nat. Rev. Neurosci.* **2003**, *4*, 165–178. [CrossRef] [PubMed]
38. Frith, C.D. The social brain. *Philos. Trans. R. Soc. Lond. B* **2007**, *362*, 671–678. [CrossRef] [PubMed]
39. Mead, G.H. *Mind, Self and Society*; University of Chicago Press: Chicago, IL, USA, 1934; Volume 111.
40. Collins, H.M. *Artificial Experts: Social Knowledge and Intelligent Machines (Inside Technology)*; MIT Press: Cambridge, MA, USA, 1990.
41. Yudkowsky, E. Levels of organization in general intelligence. In *Artificial General Intelligence*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 389–501.
42. Dodds, A.E.; Lawrence, J.A.; Valsiner, J. The Personal and the Social: Mead's Theory of the Generalized Other. *Theory Psychol.* **1997**, *7*, 483–503. [CrossRef]

43. Microsoft Research India Workshop on ASI. 2007. Available online: https://www.microsoft.com/en-us/research/event/microsoft-research-india-summer-school-artificial-social-intelligence/ (accessed on 28 June 2018).
44. Danescu-Niculescu-Mizil, C.; Gamon, M.; Dumais, S. Mark my words!: linguistic style accommodation in social media. In Proceedings of the 20th International Conference on World Wide Web, Hyderabad, India, 28 March–1 April 2011; pp. 745–754.
45. Bohus, D.; Sean, A.; Mihai, J. Rapid development of multimodal interactive systems: a demonstration of platform for situated intelligence. In Proceedings of the 19th ACM International Conference on Multimodal Interaction, Glasgow, Scotland, 13–17 November 2017; pp. 493–494.
46. Bohus, D.; Eric, H. Managing human-robot engagement with forecasts and ... um ... hesitations. In Proceedings of the 16th International Conference on Multimodal Interaction, Istanbul, Turkey, 12–16 November 2014; pp. 2–9.
47. Voss, P. From Narrow to General AI. Medium, 2017. Available online: https://medium.com/intuitionmachine/from-narrow-to-general-ai-e21b568155b9 (accessed on 4 October 2017).
48. Goertzel, B. *Artificial General Intelligence*; Pennachin, C., Ed.; Springer: New York, NY, USA, 2007; Volume 2.
49. Voss, P. Essentials of general intelligence: The direct path to artificial general intelligence. In *Artificial General Intelligence*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 131–157.
50. Wang, P.; Goertzel, B. (Eds.) *Theoretical Foundations of Artificial General Intelligence*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2012; Volume 4.
51. Ye, P.; Wang, T.; Wang, F.-Y. A Survey of Cognitive Architectures in the Past 20 Years. *IEEE Trans. Cybern.* **2018**, *99*, 1–11. [CrossRef] [PubMed]
52. Newell, A.; Simon, H.A. *GPS, a Program that Simulates Human Thought*; No. P-2257; Rand Corp: Santa Monica, CA, USA, 1961.
53. Lenat, D.B. CYC: A large-scale investment in knowledge infrastructure. *Commun. ACM* **1995**, *38*, 33–38. [CrossRef]
54. Laird, J.E.; Newell, A.; Rosenbloom, P.S. Soar: An architecture for general intelligence. *Artif. Intell.* **1987**, *33*, 1–64. [CrossRef]
55. Richard, L.L.; Newell, A.; Polk, T.A. *Toward a Soar Theory of Taking Instructions for Immediate Reasoning Tasks*; Carnegie Mellon University: Pittsburgh, PA, USA, 1989.
56. Goertzel, B.; Pennachin, C.; Senna, A.; Maia, T.; Lamacie, G. Novamente: An integrative architecture for Artificial General Intelligence. In *AAAI Symposium on Achieving Human-Level Intelligence through Integrated Systems and Research*; The AAAI Press: Menlo Park, CA, USA, 2004.
57. Goertzel, B. *The Structure of Intelligence*; Springer-Verlag: New York, NY, USA, 1993.
58. Goertzel, B. *The Evolving Mind*; Gordon and Breach: New York, NY, USA, 1993.
59. Goertzel, B. *Chaotic Logic*; Plenum Press: New York, NY, USA, 1994.
60. Goertzel, B. *From Complexity to Creativity*; Plenum Press: New York, NY, USA, 1997.
61. Goertzel, B. *Creating Internet Intelligence*; Plenum Press: New York, NY, USA, 2001.
62. Ramamurthy, U.; Baars, B.J.; D'Mello, S.K.; Franklin, S. LIDA: A working model of cognition. In *Proceedings of the 7th International Conference on Cognitive Modeling*; Fum, D., Missier, F.D., Andrea Stocco, A., Eds.; Edizioni Goliardiche: Trieste, Italy, 2006; pp. 244–249.
63. Schmidhuber, J. Gödel machines: Fully self-referential optimal universal self-improvers. In *Artificial General Intelligence*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 199–226.
64. Wang, P. Theories of Artificial Intelligence. In *Theoretical foundations of artificial general intelligence*; Wang, P., Goertzel, B., Eds.; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2012; Volume 4.
65. Perez, C.E. Conversational Cognition: A New Measure for Artificial General Intelligence. Medium. 2018. Available online: https://medium.com/intuitionmachine/conversational-cognition-a-new-approach-to-agi-95486ffe581f (accessed on 9 October 2018).
66. Dreyfus, H.L. *What Computers Can't Do: The Limits of Artificial Intelligence*; Harper & Row: New York, NY, USA, 1979; Volume 1972.
67. Heidegger, M. *Being and Time*; Macquarrie, J.; Robinson, E., Translators; Harper: New York, NY, USA, 1962.
68. O'Keefe, J.; Dostrovsky, J. The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Res.* **1971**, *34*, 171–175. [CrossRef]
69. O'Keefe, J. A review of the hippocampal place cells. *Prog. Neurobiol.* **1979**, *13*, 419–439. [CrossRef]

70. Moser, E.I.; Kropff, E.; Moser, M.-B. Place cells, grid cells, and the brain's spatial representation system. *Annu. Rev. Neurosci.* **2008**, *31*, 69–89. [CrossRef] [PubMed]
71. Hannaford, C. *Smart Moves: Why Learning is not All in Your Head*; Great River Books: Salt Lake City, UT, USA, 1995.
72. Brooks, R.A. Elephants don't play chess. *Robot. Auton. Syst.* **1990**, *6*, 3–15. [CrossRef]
73. Kennedy, J. Swarm intelligence. In *Handbook of Nature-Inspired and Innovative Computing*; Springer: Boston, MA, USA, 2006; pp. 187–219.
74. Dennett, D.C. Cog: Steps towards consciousness in robots. In *Conscious Experience*; Ferdinand Schoningh: Paderborn, Germany, 1995; pp. 471–487.
75. Brooks, R.A. The cog project. *J. Robot. Soc. Jpn.* **1997**, *15*, 968–970. [CrossRef]
76. Lakoff, G.; Johnson, M. *Philosophy in the Flesh*; Basic Books: New York, NY, USA, 1999; Volume 4.
77. Lakoff, G.; Johnson, M. *Metaphors We Live by*; The University of Chicago Press: London, UK, 2003.
78. Goffman, E. *Frame Analysis: An Essay on the Organization of Experience*; Harvard University Press: Cambridge, MA, USA, 1974.
79. Hebb, D.O. *The Organization of Behavior: A Neuropsychological Theory*; John Wiley & Sons, Inc.: New York, NY, USA, 1949.
80. Regier, T. A model of the human capacity for categorizing spatial relations. *Cogn. Linguist.* **1995**, *6*, 63–88. [CrossRef]
81. Winfield, A.F.T.; Blum, C.; Liu, W. Towards an ethical robot: internal models, consequences and ethical action selection. In *Conference towards Autonomous Robotic Systems*; Springer: Cham, Switzerland, 2014; pp. 85–96.
82. Conn, A. How Do We Align Artificial Intelligence with Human Values? Future of Life Institute. Available online: https://futureoflife.org/2017/02/03/align-artificial-intelligence-with-human-values/ (accessed on 3 September 2018).
83. Yudkowsky, E. *Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures*; The Singularity Institute for Artificial Intelligence: San Francisco, CA, USA, 2001.
84. Yudkowsky, E. *Coherent Extrapolated Volition*; Singularity Institute for Artificial Intelligence: San Francisco, CA, USA, 2004.
85. Tegmark, M. *Life 3.0: Being Human in the Age of Artificial Intelligence*; Knopf: New York, NY, USA, 2017.
86. Hardesty, L. Crowdsourced Tool for Depression. MIT News. Available online: http://news.mit.edu/2015/crowdsourced-depression-tool-0330 (accessed on 30 March 2015).
87. Postman, N. *Technopoly: The Surrender of Culture to Technology*; Vintage Books: New York, NY, USA, 1992.

information

MDPI

Article

When Robots Get Bored and Invent Team Sports: A More Suitable Test than the Turing Test?

Hugh Trenchard

Independent Researcher, V8V 1S9 Victoria, Canada; h.a.trenchard@gmail.com; Tel.: +1-250-472-0781

Received: 9 April 2018; Accepted: 8 May 2018; Published: 11 May 2018

Abstract: Increasingly, the Turing test—which is used to show that artificial intelligence has achieved human-level intelligence—is being regarded as an insufficient indicator of human-level intelligence. This essay extends arguments that embodied intelligence is required for human-level intelligence, and proposes a more suitable test for determining human-level intelligence: the invention of team sports by humanoid robots. The test is preferred because team sport activity is easily identified, uniquely human, and is suggested to emerge in basic, controllable conditions. To expect humanoid robots to self-organize, or invent, team sport as a function of human-level artificial intelligence, the following necessary conditions are proposed: humanoid robots must have the capacity to participate in cooperative-competitive interactions, instilled by algorithms for resource acquisition; they must possess or acquire sufficient stores of energetic resources that permit leisure time, thus reducing competition for scarce resources and increasing cooperative tendencies; and they must possess a heterogeneous range of energetic capacities. When present, these factors allow robot collectives to spontaneously invent team sport activities and thereby demonstrate one fundamental indicator of human-level intelligence.

Keywords: artificial intelligence; Turing test; embodiment; competition; cooperation; self-organization; robots; heterogeneity; team sports

1. Introduction

In his book, “The singularity is near: when humans transcend biology” [1], Ray Kurzweil predicts that human-level artificial intelligence will be achieved roughly in the year 2029. Then, Kurzweil argues, computers will be able to pass the Turing test. The Turing test is met when humans, in conversation with artificial intelligence, will not be able to distinguish the artificially intelligent machine from a human [2]. This, says Kurzweil, will lead to a fundamental shift in mind, society and economics in about 2045—an event which he refers to as the singularity.

One criticism of the Turing test and Kurzweil’s prediction is that they focus too much on the human brain as the source of the measurable computational power of human intelligence, while largely ignoring or minimizing the influence of the human body on human intelligence. In this view, it is a mistake to measure human intelligence as a function of the computational capacity of the brain in the absence of a functional body: the totality of human intelligence can only be properly measured and simulated by locating any artificial intelligence within a functional body.

Similarly, the Turing test is increasingly viewed as insufficient for proving that artificial intelligence has achieved human-level intelligence [3,4]; the test has been criticized for focusing on problems of language and reasoning, while problems of motor skills and perception are absent [5]. Basic physical motor and perceptual skills require substantially greater computational capacity than tasks typically associated with higher-level intelligence such as playing checkers or solving problems on intelligence tests. This describes Moravec’s paradox [6] (p. 15) [7] (p. 29), and implies that the Turing test does not encompass even basic levels of individually expressed human intelligence.

One may conclude, therefore, that embodied humanoid robot intelligence is the appropriate analog of human-level artificial intelligence, and not merely computer computational power activated to solve abstract problems in the absence of a functional body.

In response to this view, Kurzweil says [1] (p. 260):

> " . . . the real issue involved here is strong AI (artificial intelligence that exceeds human intelligence). The standard reason for emphasizing robotics in this formulation is that intelligence needs an embodiment, a physical presence to affect the world. I disagree with the emphasis on physical presence, however, for I believe that the central concern is intelligence. Intelligence will inherently find a way to influence the world, including creating its own means for embodiment and physical manipulation."

Here Kurzweil rejects a critical conceptual aspect of embodiment: that human intelligence arises largely when humans exploit their physical environments and modify their behavior according to the ecological niche in which they interact collectively and compete for resources [8]. Contrary to Kurzweil's argument, it is not enough evenfor humanoid robots to attain sufficient intelligence and the sensorimotor capacity required to engage in human-like physical activities, such as to play a team-sport like soccer. Indeed, it would be self-evident that humanoid robots playing soccer at human levels would possess sufficient computational power for many of the high-capacity sensorimotor functions that are indicated by Moravec's paradox. Further, Kurzweil addresses any shortcomings in computational capacity that may be required for brain processes otherwise unaccounted for by his current projections:he posits that even if the computational capacity of computers is off by a factor of a billion, this would delay the singularity by only 21 years [1] (p. 123).

However, the problem is not strictly one of computational capacity, since massive computational capacity does not seem to guarantee human-level intelligence. Instead, the problem relates to the uniquely human traits of collective intelligence and self-organized behavior, which emerge in appropriate conditions and simultaneously drive the development of human intelligence, and clearly indicate that such intelligence is present.

Thus, rather than merely possessing sufficient capacity at the individual level to engage in uniquely human collective behaviors, for humanoid robots to establish human-level intelligence, they must demonstrate that they can self-organize human-level behaviors in appropriate ecological conditions. If these, or similar, human-like collective behaviors do not self-organize, then the test for human-level intelligence has not been met. Here we make the case for one such test: when robots self-organize or invent team sports. For a comprehensive definition and review of self-organizing systems, see [9].

Necessarily, such an invention will occur only after humanoid robots are programmed with the basic sensorimotor capacity to play team sports like soccer, because the converse cannot be true; i.e., robots must first be physically capable of playing soccer before they can invent a game in which they exploit those capacities. Since robots are projected to acquire the capacity to play soccer by 2050 [10]—obviously well beyond 2029—we may conclude that robots will not invent the game until sometime after 2050, and therefore robots will not pass the "team-sport test" in the near future. This does not mean that robots will never invent team sports; the key point is that the Turing test is inadequate for proving human-level intelligence, and that further benchmarks of robot intelligence are required to establish true human-level intelligence.

In this paper, we discuss what is meant by the "invention" of team sport and identify the necessary conditions for this to occur. The proposed test may be considered a first level or minimum threshold of easily observed and recognized human-level collective intelligence. As a minimum threshold, it does not preclude other tests that demonstrate the emergence of social, economic, and evolutionary collective behaviors at similar or greater levels of complexity. Thus, similar tests may be proposed requiring a variety of self-organized, human-like social and economic activities, or certain biological dynamics such as evolutionary processes [11].

2. Defining Human-Level Intelligence

It is helpful to set a framework for intelligence in the context of embodied and self-organized collective behavior. To this end, we adopt Winfield's [12] (pp. 2–3) four components of embodied intelligence:

- Morphological intelligence—"the physical behavior that emerges from the interaction of the body, its control systems and the environment".
- Swarm intelligence—collective behavior is distributed and decentralized.
- Individual intelligence—"the ability to both respond (instinctively) to stimuli and, optionally, learn new—or adapt existing—behaviours through a process of trial and error".
- Social intelligence—"the kind of intelligence that allows animals or robots to learn from each other".

Winfield [12] (p. 6) concludes that when considering the integration of these four components of intelligence, "the intelligence of intelligent robots falls far short of that of most animals"; i.e., no single component of embodied intelligence sufficiently encompasses the breadth of animal, let alone human, intelligence.

Under this broad framework, we consider the swarm and social components of humanoid robot intelligence, and identify a threshold level of complex behavior by which we may confidently declare that robots have achieved human-level swarm and social intelligence: the emergence, or invention, of team sport.

Further, we assume that team sport inherently involves human-level intelligence since, although many animals may have a sense of competition, they are well-understood not to engage in team sport [13] as subsequently defined. Therefore, it is not necessary to consider in detail other definitions and measures of intelligence. For a review of some 70 such definitions, see [14].

3. The Embodiment of Artificial Intelligence

We do not review the concept of embodiment in detail, although a brief overview is useful. For a full treatment of the embodiment argument, see [8].

The embodiment school of artificial intelligence is perhaps traced to Brooks [15] (p. 3), who argues that four basic properties of robot functionality distinguish robot intelligence from computer intelligence and its architectural constraints:

- Situatedness—robots are located in the world.
- Embodiment—robots have bodies in which they directly experience the world.
- Intelligence—the source of intelligence derives largely from the physical coupling between the robot and the world.
- Emergence—robot intelligence emerges from interactions among its system components, and with the world.

As Pfeifer and Bongard [8] argue, of these four properties, embodied intelligence emerges from the integrated brain and body; when we evaluate intelligence, the contribution of the functioning body to overall intelligence cannot be ignored. Similar proponents of this school include Clarke [16], whose "extended mind" hypothesis expands intelligence to be fundamentally connected with the tools that humans use to solve problems. Moreover, embodied intelligence implies the dynamical evolutionary and ecological processes that have led to the existing state of the human body [8] (pp. 178–213).

Thus, even if computationally equivalent to the capacity of the human brain, computer intelligence falls short of human-level intelligence in the absence of a wider perceptual and material integration with its physical environment. This has been argued in the context of a wide range of uniquely human characteristics that cannot be expected to arise in machine intelligence, even if machines are computationally equivalent to or greater than humans [17]. Cariani [18] makes a similar argument that without embodiment, computers have no mechanism by which to experience the external world and therefore cannot adapt and learn, evolve accordingly, and develop true human-level intelligence.

Similarly, Moravec [6] observes that a billion years of sensorimotor evolution is encoded in the human brain and implicitly also encoded in the brains of many animal species: these highly evolved processes require brain power far greater than that for abstract reasoning, which Moravec describes as "the thinnest veneer of human thought" (p. 15). By extension, embodied human-level intelligence must be located in humanoid robots in order to simulate human sensorimotor skills and to be capable of solving ordinary human problems. This includes those problems inherent in sports activities that are naturally solved through functional human physiology, such as by use of hands, or through the use of tools, as argued by Clarke [16].

Thus, if we accept the premise that the constrained and inanimate hardware of a computer—though it may contain a vast computational capacity—is insufficient to achieve the intelligence of embodied human intelligence, then any search for artificial human-level intelligence shifts in focus to humanoid robotic intelligence. The question is then whether a robot, embodying an artificial brain, can achieve human-level intelligence by integrating massive computational capacity together with a functional and dynamic body that experiences, reacts, adapts to, and learns from its interactions with other robots and their environments. Further, how will we recognize this integration when it occurs? We suggest that self-organized team sport is an easily-observed and identifiable collective human behavior that clearly indicates human-level intelligence.

4. The Transition from Team-Like Behavior to Bounded Competition and the Invention of Team Sport

Here team sport is defined as a contest between groups of participants, performed within specified spatial boundaries with specific rules of play that apply equally to both or all the participating teams (the "game"). The game objective is for one team to win the contest by accumulating some agreed-upon advantage that is greater than that accumulated by the opposing side, typically achieved by scoring goals. Game rules may vary in their constraints on the physical movements of the players, but in general the rules are minimally restrictive to permit high degrees of physical freedom by which participants rely upon their collective skill, strength, and cooperative strategy to best their opponents.

We refer to this kind of game as bounded competition. This excludes dyadic competitions (one-on-one) and competitions involving multiple competitors who do not cooperate with any team-mates, such as a 100 m sprint in which all the competitors compete within their own designated lanes and there is no expectation of cooperation. These types of competition are excluded in order to isolate for our analysis those games in which there is a clear cooperative element among team-members, in addition to the clearly competitive component of a contest between teams.

Bounded competition, when it exists, should be obvious to average adult human observers, although in some cases this may not always be so. For example, competitive mass-start bicycle racing involves, at one level, obvious team distinctions because team-members are clothed in identical racing attire, which is different between teams. However, at another level, less obvious cyclist team behavior occurs when alliances form temporarily among competitors from different teams or during events when not all participants are part of specified teams. These alliances may continually re-constitute among different cyclist combinations, and observers may not easily recognize when riders are working together or in opposition. This is due to cyclists' continuous positional change within a peloton (group of cyclists), and the tactics that derive from the energy-saving advantages of cycling in positions behind competitors where power requirements are reduced. This kind of less obvious team behavior may be characterized as proto-team behavior, or team-like behavior. A further, more subtle form or characteristic of team behavior involves creative and novel pattern formations that emerge among team interactions [19].

Thus, there is an implied threshold between team-like or proto-team activity—which is part of a gradual multi-generational process originating in various social activities and may occur among many animal species—and uniquely human team sport. In this way, the invention of team sport describes a phase transition that occurs when team-like behavior becomes true team sport. In Section 6 we cite studies involving swarm proto-team behavior.

This transition may involve the simultaneous emergence of creative and novel pattern formation, as examined by Hristovski et al. [19]. Such novel pattern formation may pre-exist the invention of team sport and emerge alongside team-like behavior. However, like the amorphous team-like behavior among pelotons, the presence of team sport in these circumstances may not be obvious to the average adult human observer. Thus, one appropriate clear indication of human-level humanoid robot intelligence in this context is the emergence, or invention, of bounded competition in the form of actual team sport.

When team-sport emerges among robots as a product of self-organized processes, it is implied that robots are not asked to play sports by external human sources or programmed specifically for these processes. The self-organized process suggests that robots would develop the games themselves and find ways to agree collectively upon the rules of the game, constructed as a function of environmental constraints and individual robot kinematics. This kind of rule-formation is distinguishable from rules that are pre-programmed for the robots.

5. The Origins of Team Sport

Collective sporting activities are universal among humans [20], and are driven by underlying universal physical and physiological principles [21]. Sports in general involves complex organizational structures and dynamics that are both self-organized [22,23] and planned according to prescribed strategies and tactics.

Studies indicate several possible origins of team sport, including the impulse to play, or the need to practice hunting abilities ([24], and references therein); sport may achieve cultural objectives including the ritualistic, cultic and cathartic ([25], and references therein). There is a connection between sport and military practice and the discharge of aggressive urges, particularly among combative sports ([26], and references therein). Athletic success may also confer selective reproductive advantages [25].

Arguably, objectives such as practicing hunting abilities and military practice are more primitive than other more sophisticated cultural objectives, since hunting and war are rooted more directly in essential resource acquisition. By extracting the most primitive origins of team sport, we may infer that if robots are to invent team sport, the invention process is more likely to originate as an artifact of, or simultaneously with, basic resource or energy acquisition, rather than as an artifact of more developed ritualistic or other cultural complexities. Thus, we suggest that proto-sporting activities tend to emerge in the context of basic competitive and cooperative behaviors that are necessarily connected with the energetic requirements of the individuals involved.

6. The Emergence of Human Collective Behavior

Humans engage in complex collective activities in which novel behavior emerges, and, as posited, this kind of behavior must be observed in robots too before robots may be viewed as having achieved human-level intelligence. As Cariani [18] (p. 48) stated, "If we want our devices to be creative in any meaningful sense of the word, they must be capable of emergent behavior, of implementing functions we have not specified". Thus, embodied intelligence must interact collectively such that novel patterns of behavior emerge that are uniquely human in nature. Furthermore, if human-level embodied intelligence is to be achieved, it is reasonable to expect intelligence to emerge in typical environments or conditions in which humans operate.

As noted, there are numerous social and economic contexts and environmental conditions in which these behaviors may be observed. Perhaps the most basic and universal of such conditions is resource scarcity, and the competition for those resources. Some progress in swarm robotics has been achieved to model collective behavior among robots [27] which includes solving resource allocation problems. Such advances include: group transport of objects [28], shortest-path finding [29], task allocation [30], energy foraging [31,32], and communication-based navigation [33]. In addition, basic elements of hierarchical cooperation have been demonstrated involving team-work [34], as has coordination between types of robot swarms [35,36]. For a review of progress in robot collective behavior, see Bayandir [37].

Overall, these advances remain at a comparatively primitive level, and are no more developed than the insect collective behaviors from which they are inspired [38], though even insects' individual physical abilities currently surpass robot abilities of comparable size [37].

Given this, it seems that the invention of team sport by humanoid robots remains a distant prospect. However, it is also conceivable that robots will never invent team sport if appropriate conditions are not present, or if robots are not immersed within an appropriate physical environment that fosters the invention of team sport.

7. Necessary Conditions for the Emergence of Team-Sport

Since sports are played within arbitrary three-dimensional boundaries, an obvious necessary condition for sport is that participants exhibit some minimal level of kinematic agility to interact in three-dimensional space. On the other hand, one could argue that computers, absent of any kinematic functions, may well self-organize cooperative-competitive dynamics akin to that of team sports in a virtual environment within the confines of their hardware and across networks. Perhaps such games could involve clusters of networks ("teams") and network boundaries (the "field of play") that are arbitrarily agreed upon by thrill-seeking computers. As a possible precursor to this, advances are currently underway in computer-human collaborative networks [39].

However, even if this is possible, unless architecturally-constrained computers develop a way to reveal their game playing activities, it will be difficult from a human perspective to observe and record computer sports that are played strictly in abstract cyberspace. Also, any evidence of such virtual game playing may be difficult to extract. Further, unless computers engage in some form of resource competition in the first place—which although not inconceivable, would perhaps be difficult to find evidence for—it seems unlikely computers will invent game playing as an artifact of real-life competition for resources.

In addition to a minimal level of robot agility, the following necessary conditions are proposed as requisite for the emergence of robot team sport, although other conditions undoubtedly exist:

- The intrinsic capacity of humanoid robots to compete and cooperate for resources.
- Sufficient periods of leisure time during which robots engage in simulated or artificial resource gathering activities that represent a form of proto-team sport, leading to an eventual transition to actual team sport.
- Heterogeneous robot energetic capacities.

7.1. The Capacity of Humanoid Robots to Compete and Cooperate for Resources

Biological organisms compete and cooperate for various kinds of resources [40] the most basic of which include food, water, and protection from harmful elements. Presently, humans supply these resources to computers in the form of electricity and hardware, and computers do not compete for these resources. Although computers by themselves do not generally possess physical mechanisms to install more hardware or to connect themselves to electrical outlets, robots certainly can be designed with these abilities and the capacity to forage for resources without requiring humans for resource inputs.

Robot science is currently sufficiently advanced for primitive levels of cooperative resource gathering [31], as noted. Foraging robots also participate in limited competition when more than one robot seeks the same energy source, except that the competitive response is benign and there appears to be no existing algorithm that causes competing robots to modify their movements toward resources to acquire it first or to fight for the resource [32].

It does not require much imagination to consider the consequences of more advanced competition among robots to acquire resources, and undoubtedly robots may be programmed to engage in fight and flight responses to acquire resources ahead of competitors. This by itself poses a challenge for robot ethicists and policy-makers, and arguably robots need not be programmed with these competitive capacities.

However, it is suggested that such a capacity is indeed a necessary pre-condition to the emergence of team sport, in addition to some intrinsic robot capacity to cooperate. Thus, without a natural propensity to compete and cooperate for resources, it is unlikely that team sport will emerge among robots as an artifact of that process.

7.2. Leisure Time as a Necessary Condition for the Emergence of Robot Team Sports

Broadly, in times of scarcity, people invest disproportionately greater durations of their time and energy in activities for basic survival. Similarly, when people experience elevated degrees of scarcity, they must conserve resources and cannot spare the high energetic costs of sporting activity. Thus, in conditions of great scarcity, it is reasonable to conclude that there is insufficient leisure time for organized sports to develop.

This connection between sports participation and relative affluence is revealed in literature from the United States and the United Kingdom [41]. Similarly, research has shown that having children reduces parents' participation in sport [41,42] presumably because parents' energy and time resources are exhausted during the process of child-rearing. Eberth and Smith [43] found that parents with children under two years old had negatively impacted participation in sports, while parents with children between ages two and 15, did not. Ruseski et al. [42] concluded that individuals with a higher income were more likely to participate in physical activity in general, but may spend less time engaged in that activity.

Downward and Riordan [41] suggest that in some cases, the form of employment and level of education are more influential than work hours and household income levels in determining sports participation. For instance, the authors showed that increased education favors increased income, but this results in work-time constraints on sport participation, and that sport participation tends to increase when time spent working is reduced and flexible. By contrast, the authors show that reductions in sport participation increase with age and with being the individual responsible for housekeeping; similar reductions in participation are shown when people undertake voluntary work or are semi-skilled.

Clearly, sports participation is a complex amalgam of variable social, cultural, and economic factors. Nonetheless, these factors indicate that sport participation is largely traceable to increased leisure time and energetic abundance. Considering a primitive socio-economic setting nearer to the origins of human species in evolutionary development, the key drivers of sports participation appear closely linked with abundant time and energy.

So, if humanoid robots are expected to invent team sport spontaneously as a function of their innate intelligence and without being programmed to do so, their resources cannot be entirely allocated toward resource gathering and other activities, and they must have sufficient energy to engage in the high-energy physical activities inherent in team sport. Put less formally, robots must be permitted "down-time" and be allowed to become sufficiently bored to figure out how else they might want to use their time and energy. Robots with human-level intelligence will not sit idly by.

7.3. The Heterogeneity Requirement and the Energetic Threshold for the Emergence of Team Sport

If team sports are largely traceable to primitive conditions for resource competition and cooperation, and tend to emerge in times of relative energetic abundance, what is the collective energetic threshold at which we might expect human-like cooperative-competitive activities to emerge and give rise to the invention of team sports?

To illustrate the threshold of energetic abundance or degree of leisure time at which team sport might emerge, we apply principles of bicycle peloton behavior. As previously stated, a peloton is a group of cyclists who, by riding in zones behind others, save energy by drafting. Cyclists at the front of the group encounter the highest energetic costs because they directly face the wind. Cyclists' intrinsic abilities are heterogeneous and tend to span a narrow range [44,45].

In simulated pelotons [44], a threshold between a predominantly collective competitive state and a predominantly collective cooperative state occurs when weaker cyclists can sustain the pace set by the strongest cyclists only by exploiting energy-saving drafting positions. When the pace set by leading cyclists is too high for weaker cyclists to move to the front and to share the costliest front positions, cyclists are incapable of cooperating with other cyclists. In this state, cyclists are predominantly competitive as a matter of physiological necessity, during which time cyclists operate as "every man for himself".

When the pace set by leading cyclists is sufficiently low, weaker cyclists can advance to the head of the peloton, and thereby cooperate by sharing the costliest front positions. In this state, cyclists are predominantly cooperative.

The threshold between the predominant competitive and cooperative states may be quantified as the difference between the power output of the pace-setting cyclist and the maximum sustainable power of the drafting cyclist, when that difference does not exceed the magnitude of energy saved by drafting [44]. In other words, a weaker cyclist can sustain the pace of a stronger one if she is no weaker than the equivalent reductions in power requirements afforded by drafting.

The presence of this threshold implies that if cyclists were entirely homogeneous in capacity, their collective state would be a constantly cooperative one in which they could always share the costliest positions, with no competitive tension. Conversely, if the difference between cyclists' individual capacities were too high, cyclists would proceed in a constantly competitive state in isolation from each other. To illustrate the latter, consider the extreme differences in the energetic capacities of an ant and a bird in flight—the two simply cannot cooperate and each must fend for itself in isolation from the other.

By implication, the principle also applies to variability in expended energy: cyclists could be initially homogeneous in energetic or metabolic capacity, but expend their energetic resources asymmetrically. For instance, cyclists who spend a lot time facing the wind expend far more energy than those who occupy drafting positions, even if they are all equally strong. In effect, this causes a variable range of cyclists' output capacities, thus increasing the likelihood for the emergence of the competitive phase.

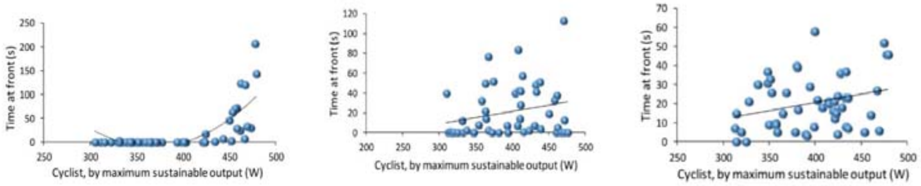
Peloton simulations demonstrate this principle; when peloton speeds are comparatively low, cyclists collectively tend to share the costliest front positions randomly, whereas when peloton speeds are high and cyclists approach their sustainable power thresholds, front positions become dominated by stronger cyclists, as shown in Figure 1 [44].

The sharing of costly positions at low levels of energy expenditure leads to a general hypothesis that cooperation among heterogeneous agents tends to emerge in times of relative abundance; i.e., when individuals' energetic resources are not maximally expended or strained. This is consistent with literature which suggests that participation in sports tends to occur during times of leisure or energetic abundance, as discussed.

We propose that this differential or heterogeneous range in energetic capacities, whether intrinsic to robots or generated over time through asymmetric energy consumption, is a necessary pre-condition to the emergence of the competitive-cooperative tension that is foundational to the invention of team sports.

In terms of collective humanoid robots, without data for the robots' range of energetic capacities, the types of tasks they will undertake, and the energy required for those tasks, it is not possible to suggest where this collective competitive-cooperative threshold may lie. Nonetheless, we argue in principle that for robots to spontaneously invent team sports, their collective outputs must be below this threshold such that robots enjoy comparatively low energetic expenditure in a predominantly cooperative state.

Studies of heterogeneous robot swarms are a relatively recent area of robot study since, in past, the field has tended to focus on homogeneous swarms [46–48]. Ranjbar-Sahraeia et al. [49] distinguish between "soft" heterogeneity in robotics as describing the range of individual differences among a collective, and "hard" heterogeneity as differences in type. As the peloton analogy suggests, such soft differential energetic capacities among collective members are natural and foundational to the variations in competitive and cooperative drive that emerges as a function of resource scarcity. We may predict, therefore, that team sports are likely to emerge only among heterogeneous robot energetic capacities.



(**a**) (**b**) (**c**)

Figure 1. (**a**) Simulated cyclists travel at high speeds relative to their maximal capacities. The peloton self-organizes so that strongest simulated cyclists tend to spend the most time in highest cost front positions (R = 0.7547); (**b**) Simulated cyclists travel at medium speeds relative to their maximal capacities. Here there is no discernable preference for cyclists of any strength to dominate the front position. While they do not all share the highest cost position equally, sharing is more randomly distributed (R = 0.2825); (**c**) Simulated cyclists travel at low speeds relative to their maximal capacities. Like cyclists traveling at mid-speeds, there is no discernable preference for cyclists of any strength to dominate the front position (R = 0.2546). Image adapted from [44] (Figures 4–6).

8. The Status of Robo-Soccer

Soccer is an obvious choice of team sport by which to test the skills of robots due to its accessibility and universality [50], and wide range of required physical skills including running at variable speeds, frequent directional changes, jumping, sliding, ball-handling with all parts of the body (including arms for throw-ins and for goalies), and highly coordinated team play.

The challenge of achieving robot team sport activity is well-recognized among the artificial intelligence community; to this end robot soccer events have been an annual international event since 1997 [51], and perhaps first formally proposed by Sahota and Mackworth [52].

The goal for researchers is to produce a robot team that can beat the best human team, which is thought to be achievable by 2050 [10]. As of 2015, there are many existing challenges to achieving this goal, including: robot running, high-kicks, jumping, controlled landings, walking/running over uneven surfaces, ball throwing (done by goalies, and from sidelines after the ball has gone out of play), ball receiving, in addition to integrated team organization [53]. Despite the currently primitive state of robot soccer, there is optimism that even by 2030, robots will be technically capable of playing against an "unprofessional human team" [53]. For a video of a 2017 humanoid robot soccer competition, see [54] and for an extended video that includes wheeled robots, see [55].

For robot soccer, robots are pre-programmed with either autonomous or centralized control. Autonomous robots are mutually independent with their own instructions about how to respond in given situations and based on local information gleaned from within robots' field of view; centrally controlled robots respond based on a single strategy that applies to all robots according to globally-available information about the positions of all the other players [56].

For humanoid robots to master soccer is a clearly a monumental achievement, which makes great strides toward proving that embodied humanoid artificial intelligence is equivalent to human intelligence. As suggested, however, to establish equivalence between humanoid robot intelligence and human intelligence, it is not enough that robots are demonstrably capable of playing soccer at human levels. A further critical step in humanoid robot intelligence must exist alongside the physical mastery of soccer: humanoid robots must spontaneously invent the game of soccer, or similar games or team sports. Robots must decide, independently of human instructions, to make their own games that involve robot competition and cooperation. To show that robots are capable of playing soccer is merely a preceding first step to establish that if robots did invent soccer, they could in fact play it. Since human intelligence includes a predisposition to invent soccer, not just to play it, it is the invention of soccer, or its spontaneous self-organization, that is a critical test for robot intelligence.

9. Conclusions

Although computer computational capacity may well soon match and exceed the computational capacity of the human brain, we have argued that for artificial intelligence to match human intelligence, it must first be embodied. Secondly, it must develop collective behaviors through similar processes and in similar conditions as those by which humans have evolved. The basic necessary conditions include: variable resource scarcity and abundance during which robots develop an inherent propensity to compete and to cooperate; sufficient periods of leisure time in which robots may, without being programmed to do so, spontaneously invent team sports; and a heterogeneous range of robot energetic capacities.

Given these factors, and that the benchmarks for human intelligence have not yet been exhausted, a novel test is proposed for artificial intelligence to prove it has achieved human-level intelligence: the invention of team-sport test. In seeking to extend the test for artificial intelligence in this way, one may ask, is it not enough that robots acquire the computational capacity required to engage the myriad sensorimotor skills to play soccer at a human level, which is predicted to be achieved by 2050? Why shift the goal-posts one step farther and demand that robots not only demonstrate the ability to play team sports, but that robots must then invent team-sports?

Indeed, Kurzweil [1] (p. 292) has argued that "as long as there are discrepancies between human and machine performance—areas in which humans outperform machines—strong AI skeptics will seize on these differences." Kurzweil [1] (p. 290) lists many examples of functions once thought to be only the domain of humans as now within the capacity of computers, including: diagnosing electrocardiograms, composing in the style of Bach, recognizing faces, guiding missiles, playing ping-pong and mastering chess, picking stocks, improvising jazz, proving important theorems, and understanding continuous speech.

Further, even if we accept a new test for the emergence of self-organized team-sports, is there any reason to believe the test will not be passed soon after Kurzweil's predicted singularity if not contemporaneously to it? When we assess the comparatively primitive state of humanoid robots and their cooperative dynamics generally to the projected timeline for high-level robot soccer to occur around the year 2050, and a subsequent period of unknown duration to determine whether robots can invent team sport, it appears unlikely that robot intelligence will match human intelligence until well after 2050.

Still, it is impossible to deny the rapidity of technological advances, as Kurzweil [57] (pp. 30, 32) asserts, and so we may well expect that robots will indeed pass this test not long after robots acquire the ability to play human-level soccer, projected to occur by 2050. Yet the ethical question remains: is it desirable for humans to create the conditions in which robots have the capacity and the need to compete for scarce resources, given the undesirable implications of such a capacity? In one sense, the question is absurd, for it is easy to imagine that even if human robot makers control optimal robot energy levels without inducing robot competition and cooperation, eventually such control will be lost. On the other hand, at present, it remains possible for humans to control the conditions under which robots may self-organize team sports, simply by tuning the degrees of available resource abundance and robot heterogeneity. Under such present control, it remains possible that the team-sport test will never be achieved.

Acknowledgments: The author acknowledges the comments of three anonymous reviewers which helped to improve this paper, as well as the support of the editors.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Kurzweil, R. *The Singularity is Near: When Humans Transcend Biology*; Penguin Group: New York, NY, USA, 2005.
2. Turing, A.M. Computing machinery and intelligence. *Mind* **1950**, *49*, 433–460. [CrossRef]
3. You, J. Beyond the Turing Test. *Science* **2015**, *347*, 116. [CrossRef] [PubMed]
4. Grosz, B. What question would Turing pose today? *AI Mag.* **2012**, *33*, 73–81. [CrossRef]
5. Ortiz, C. Why we need a physically embodied Turing test and what it might look like. *AI Mag.* **2016**, *37*, 55–62. [CrossRef]
6. Moravec, M. *Mind Children: The Future of Robot and Human Intelligence*; Harvard University Press: Cambridge, MA, USA, 2005.
7. Minsky, M. *The Society of Mind*; Simon & Schuster: New York, NY, USA, 1986.
8. Pfeifer, R.; Bongard, J. How the Body Shapes the Way We Think—A New View of Intelligence. In *A Bradford Book*; The MIT Press: Cambridge, MA, USA; London, UK, 2007.
9. Gershenson, C.; Trianni, V.; Werfel, J.; Sayama, H. Self-Organization and Artificial Life: A Review. In Proceedings of the 2018 Conference on Artificial Life, Tokyo, Japan, 23–27 July 2018 (submitted).
10. Kitano, H.; Asada, M. The RoboCup humanoid challenge as the millennium challenge for advanced robotics. *Adv. Robot.* **1998**, *13*, 723–736. [CrossRef]
11. Nolfi, S.; Floreano, D. *Evolutionary Robotics: The Biology, Intelligence, and Technology of Self-Organizing Machines*; MIT Press: Cambridge, MA, USA, 2000.
12. Winfield, A. How intelligent is your intelligent robot? *arXiv*, 2017.
13. Do Animals Have a Sense of Competition? 2018. Available online: https://gizmodo.com/do-animals-have-a-sense-of-competition-1823122780 (accessed on 29 April 2018).
14. Legg, S.; Hutter, M. A collection of definitions of intelligence. *Front. Artif. Intell. Appl.* **2007**, *157*, 17.
15. Brooks, R.A.; Stein, L.A. Building brains for bodies. *Auton. Robots* **1994**, *1*, 7–25. [CrossRef]
16. Clark, A. *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*; Oxford University Press: Oxford, UK, 2008.
17. Braga, A.; Logan, R.K. The Emperor of Strong AI Has No Clothes: Limits to Artificial Intelligence. *Information* **2017**, *8*, 156. [CrossRef]
18. Cariani, P.A. On the Design of Devices with Emergent Semantic Functions. Ph.D. Thesis, State University of New York at Binghamton, New York, NY, USA, 1989.
19. Hristovski, R. Constraints-induced emergence of functional novelty in complex neurobiological systems: A basis for creativity in sport. *Nonlinear Dyn. Psychol. Life Sci.* **2011**, *15*, 175–206.
20. Brown, D.E. *Human Universals*; Temple University Press: Philadelphia, PA, USA, 1991.
21. Glazier, P.S. Towards a grand unified theory of sports performance. *Hum. Mov. Sci.* **2017**, *56*, 184–189. [CrossRef] [PubMed]
22. Araújo, D.; Davids, K. Team synergies in sport: Theory and measures. *Front. Psychol.* **2016**, *7*, 1449. [CrossRef] [PubMed]
23. Balagué, N.; Torrents, C.; Hristovski, R.; Kelso, J.A. Sport science integration: An evolutionary synthesis. *Eur. J. Sport Sci.* **2017**, *17*, 51–62. [CrossRef] [PubMed]
24. Scambler, G. *Sport and Society: History, Power and Culture*; Open University Press, McGraw-Hill Education: Berkshire, UK, 2005.
25. Lombardo, M.P. On the evolution of sport. *Evolut. Psychol.* **2012**, *10*. [CrossRef]
26. Sipes, R.G. War, sports and aggression: An empirical test of two rival theories. *Am. Anthropol.* **1973**, *75*, 64–86. [CrossRef]

27. Trianni, V.; Dorigo, M. Self-organisation and communication in groups of simulated and physical robots. *Biol. Cybern.* **2006**, *95*, 213–231. [CrossRef] [PubMed]
28. Gross, R.; Dorigo, M. Towards group transport by swarms of robots. *Int. J. Bio-Inspired Comput.* **2009**, *1*, 1–3. [CrossRef]
29. Sperati, V.; Trianni, V.; Nolfi, S. Self-organised path formation in a swarm of robots. *Swarm Intell.* **2011**, *5*, 97–119. [CrossRef]
30. Krieger, M.J.; Billeter, J.B.; Keller, L. Ant-like task allocation and recruitment in cooperative robots. *Nature* **2000**, *406*, 992. [CrossRef] [PubMed]
31. Zedadra, O.; Seridi, H.; Jouandeau, N.; Fortino, G. Energy expenditure in multi-agent foraging: An empirical analysis. In Proceedings of the Federated Conference on Computer Science and Information Systems (FedCSIS) IEEE, Łódź, Poland, 13–16 September 2015; pp. 1773–1778.
32. Liu, W.; Winfield, A.F. Modeling and optimization of adaptive foraging in swarm robotic systems. *Int. J. Robot. Res.* **2010**, *29*, 1743–1760. [CrossRef]
33. Ducatelle, F.; Di Caro, G.A.; Pinciroli, C.; Mondada, F.; Gambardella, L. Communication assisted navigation in robotic swarms: Self-organization and cooperation. In Proceedings of the 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), San Francisco, CA, USA, 25–30 September 2011; pp. 4981–4988.
34. Nouyan, S.; Groß, R.; Bonani, M.; Mondada, F.; Dorigo, M. Teamwork in self-organized robot colonies. *IEEE Trans. Evolut. Comput.* **2009**, *13*, 695–711. [CrossRef]
35. Ducatelle, F.; Di Caro, G.A.; Pinciroli, C.; Gambardella, L.M. Self-organized cooperation between robotic swarms. *Swarm Intell.* **2011**, *5*, 73. [CrossRef]
36. Dorigo, M.; Floreano, D.; Gambardella, L.M.; Mondada, F.; Nolfi, S.; Baaboura, T.; Birattari, M.; Bonani, M.; Brambilla, M.; Brutschy, A.; et al. Swarmanoid: A novel concept for the study of heterogeneous robotic swarms. *IEEE Robot. Autom. Mag.* **2013**, *20*, 60–71. [CrossRef]
37. Bayındır, L. A review of swarm robotics tasks. *Neurocomputing* **2016**, *172*, 292–321. [CrossRef]
38. Mohan, Y.; Ponnambalam, S. An Extensive Review of Research in Swarm Robotics. In Proceedings of the World Congress on Nature & Biologically Inspired Computing, Coimbatore, India, 9–11 December 2009.
39. Grosz, B.J. A multi-agent systems Turing challenge. In Proceedings of the 2013 International Conference on Autonomous Agents and Multi-Agent Systems, St. Paul, MN, USA, 6–10 May 2013.
40. Nowak, M.A. Five rules for the evolution of cooperation. *Science* **2006**, *314*, 1560–1563. [CrossRef] [PubMed]
41. Downward, P.; Riordan, J. Social interactions and the demand for sport: An economic analysis. *Contemp. Econ. Policy* **2007**, *25*, 518–537. [CrossRef]
42. Ruseski, J.E.; Humphreys, B.R.; Hallmann, K.; Breuer, C. Family structure, time constraints, and sport participation. *Eur. Rev. Aging Phys. Act.* **2011**, *8*, 57–66. [CrossRef]
43. Eberth, B.; Smith, M.D. Modelling the participation decision and duration of sporting activity in Scotland. *Econ. Model.* **2010**, *27*, 822–834. [CrossRef] [PubMed]
44. Trenchard, H. The peloton superorganism and protocooperative behavior. *Appl. Math. Comput.* **2015**, *270*, 179–192. [CrossRef]
45. Trenchard, H.; Ratamero, E.; Richardson, A.; Perc, M. A deceleration model for bicycle peloton dynamics and group sorting. *Appl. Math. Comput.* **2015**, *251*, 24–34. [CrossRef]
46. Szwaykowska, K.; Romero, L.M.; Schwartz, I.B. Collective motions of heterogeneous swarms. *IEEE Trans. Autom. Sci. Eng.* **2015**, *12*, 810–818. [CrossRef]
47. Gomes, J.; Mariano, P.; Christensen, A.L. Challenges in cooperative coevolution of physically heterogeneous robot teams. *Nat. Comput.* **2016**, 1–18. [CrossRef]
48. Yang, J.; Liu, Y.; Wu, Z.; Yao, M. The evolution of cooperative behaviours in physically heterogeneous multi-robot systems. *Int. J. Adv. Robot. Syst.* **2012**, *9*, 253. [CrossRef]
49. Ranjbar-Sahraeia, B.; Alersa, S.; Stankováa, K.; Tuylsab, K.; Weissa, G. Toward Soft Heterogeneity in Robotic Swarms. In Proceedings of the 25th Benelux Conference on Artificial Intelligence (BNAIC), Delft, The Netherlands, 7–8 November 2013; pp. 384–385.
50. Orejan, J. *Football/Soccer: History and Tactics*; McFarland: Glasgow, UK, 2011.
51. Osawa, E.; Kitano, H.; Asada, M.; Kuniyoshi, Y.; Noda, I. RoboCup: The robot world cup initiative. In Proceedings of the Second International Conference on Multi-Agent Systems (ICMAS-1996), Kyoto, Japan, 9–13 December 1996; AAAI Press: Menlo Park, CA, USA, 1996.

52. Sahota, M.K.; Mackworth, A.K. Can situated robots play soccer? In Proceedings of the 10th Biennial Conference of the Canadian Society for Computational Studies of Intelligence, Banff, AB, Canada, 16–20 May 1994; pp. 249–254.
53. Gerndt, R.; Seifert, D.; Baltes, J.H.; Sadeghnejad, S.; Behnke, S. Humanoid robots in soccer: Robots versus humans in RoboCup 2050. *IEEE Robot. Autom. Mag.* **2015**, *22*, 147–154. [CrossRef]
54. World Championship 2017 SPL Finals B-Human vs. Nao-Team HTWK. 2017. Available online: https://www.youtube.com/watch?v=4uYN_3gL4_YRoboSoccer (accessed on 5 April 2018).
55. RoboCup 2017. Available online: https://www.youtube.com/watch?time_continue=24395&v=BUxqFlrvkQk (accessed on 5 April 2018).
56. Snášel, V.; Svatoň, V.; Martinovič, J.; Abraham, A. Optimization of Rules Selection for Robot Soccer Strategies. *Int. J. Adv. Robot. Syst.* **2014**, *11*, 13. [CrossRef]
57. Kurzweil, R. *The Age of Spiritual Machines*; Viking: New York, NY, USA, 1999.

information

Article

Artificial Intelligence and the Limitations of Information

Paul Walton

Capgemini UK, Forge End, Woking, Surrey GU21 6DB, UK; paulnicholaswalton@gmail.com;
Tel.: +44-13-0688-3140

Received: 17 November 2018; Accepted: 18 December 2018; Published: 19 December 2018

Abstract: Artificial intelligence (AI) and machine learning promise to make major changes to the relationship of people and organizations with technology and information. However, as with any form of information processing, they are subject to the limitations of information linked to the way in which information evolves in information ecosystems. These limitations are caused by the combinatorial challenges associated with information processing, and by the tradeoffs driven by selection pressures. Analysis of the limitations explains some current difficulties with AI and machine learning and identifies the principles required to resolve the limitations when implementing AI and machine learning in organizations. Applying the same type of analysis to artificial general intelligence (AGI) highlights some key theoretical difficulties and gives some indications about the challenges of resolving them.

Keywords: information; philosophy of information; artificial intelligence; machine learning; information quality; information friction

1. Introduction

The role of artificial intelligence (AI) and machine learning in organizations and society is of critical importance. From their role in the potential singularity (for example, see [1,2]) through their more pragmatic role in day-to-day life and businesses and on to deeper philosophical questions [3] they promise to make a widespread impact on our lives. Yet, on the other hand, they are just different forms of processing information.

However, information and information processing is beset with limitations that humans do not easily notice. As Kahneman [4] says with respect to our automatic responses (what he calls System 1): "System 1 is radically insensitive to both the quality and quantity of information that gives rise to impressions and intuitions." Yet information quality, what Kahneman says we are prone to ignore, is at the heart of many fundamental questions about information. Truth, meaning, and inference are expressed using information, so it is important to understand how the limitations apply. These topics are discussed in general in [5] and in [6–8] in respect of truth, meaning and inference more particularly.

In this paper, we take the same approach to AI and machine learning and consider the questions: how do the limitations and problems associated with information relate to AI and machine learning and how can an information-centric view help us to overcome the limitations? This analysis explains some current issues and indicates implementation principles required to resolve both pragmatic and deeper issues (A note on terminology: since machine learning is a subset of AI, where the context is broad, we will refer to AI and where the context is specifically about machine learning we will refer to machine learning).

The limitations of information arise from its evolution in information ecosystems in response to selection pressures [5] and the need to make tradeoffs to tackle the underlying combinatorial and pragmatic difficulties. Information ecosystems have different conventions for managing and processing information. Think of the differences between mathematicians, banking systems and finance specialists, for example; each has their own ways of sharing information, often inaccessible to those outside the ecosystem. This approach to information is described in Section 2 that also describes the relationship of information with the interactions of Interacting Entities (IEs)—the entities, such as people, computer systems, organizations and animals that interact using information.

Following current ideas in technology architecture [9] and in usage traceable back to Darwin [10] we use the term fitness as a measure of how effectively an IE can achieve favorable outcomes in its environment. This interaction-led view leads to the following three levels of fitness that IEs may develop:

- Narrow fitness: that associated with a single interaction (and this is the type of fitness analyzed in [5–8]);
- Broad fitness: that associated with multiple interactions (of the same or different types) and the consequent need to manage and prioritize resources between the different types—this is the type of fitness linked to specialization, for example;
- Adaptiveness: that associated with environment change and the consequent need to adapt—this is the type of fitness that has led organizations to undertake digital transformation activities [11].

It is helpful to discuss fitness using some ideas developed for technology architecture [12]. Fitness needs a set of capabilities (where a capability is the ability to do something) that are provided by a set of physical components. Different components (e.g., web sites, enterprise applications, virtual assistants) are integrated together in component patterns (where the word ”pattern” is used in the sense of the technology community [13]). Just as in technology architecture, these component patterns enable or constrain the different levels of fitness.

Using this approach, Section 3 builds on the analysis in [5–8] to highlight the limitations of information, how they apply to fitness in general, how they apply to AI and how AI can help to improve fitness. This section deals with current issues with machine learning and demonstrates a theoretical basis for implementation principles to:

- Understand the levels of fitness required and their relationship with information measures (like quality, friction, and pace [5,6]);
- Analyze the integration challenges of different AI approaches—the requirements for delivering reliable outcomes from a range of disparate components reflecting the conventions of different information ecosystems;
- Understand the best way to manage ecosystems boundaries—initially, how AI and people can work together but increasingly how AI can support effective interaction across other ecosystem boundaries;
- Provide assurance about the impact and risks as AI becomes more prevalent and the issues discussed above become more important to organizational success.

The theoretical difficulties become more profound when we consider artificial general intelligence (AGI) in Section 4. The following questions highlight important theoretical difficulties for which AGI research will require good answers:

- How is fitness for AGI determined?
- How will AGI handle the integration of components, the need to accommodate different ecosystem conventions and be sufficiently adaptive?
- How will AGI process and relate abstractions and will it be able to avoid the difficulties that humans have with the relationship between abstractions and information quality?

When we analyze these questions, it is clear that there are difficult information theoretic problems to be overcome on the route to the successful implementation of AGI.

2. Selection and Fitness

The relationship between information and ideas about evolution and ecology has been studied by several authors (see for example [14,15]). This section sets out the approach to information and evolution contained in [5–8]. In this approach, information corresponds to relationships between sets of physical properties encoded using conventions that evolve in information ecosystems. Consider the elements of this statement in turn.

Information processing entities interact with their environment, so we call them Interacting Entities (IEs—people, animals, organizations, parts of organizations, political parties, and computer systems are all IEs, for example). Through interaction, IEs gain access to resources such as money, food, drink, or votes for themselves or related IEs. Through a range of processes and feedback mechanisms, derived IEs (e.g., children, new product versions, changed organizations) are created from IEs. The health of an IE—its ability to continue to interact and achieve favorable outcomes—and the nature of any derived IE depend on the resources the IE has access to (either directly or through related IEs) and the outcomes it achieves. The interactions and outcomes available, together with the competition to achieve the outcomes, define the selection pressures for any IE. The selection pressures affect the characteristics of derived IEs. Selection, in this sense, is just the result of interactions. Examples of selection pressures include the market, natural selection, elections, personal choice, cultural norms in societies and sexual selection and for any IE different combinations of selection pressures may apply.

The ability of an IE to achieve a favorable outcome from an environment state requires information processing. For any environment state an IE needs to know how to respond, so it needs to connect environment states with potential outcomes and the actions required to help create the outcomes. Thus, IEs sense the values of properties in the environment, interpret them, make inferences, and create instructions to act. This information processing results in what is sometimes called descriptive, predictive and prescriptive information [7,8], corresponding to the categorization in Floridi [16] (Please note that these terms encompass other terms for types of information, such as "knowledge" and "intelligence").

The degree to which an IE can achieve favorable outcomes we call fitness, based on the extension of the Darwin's idea [10] in modern technology development [9]. There are three levels of fitness:

- narrow fitness: the ability to achieve favorable outcomes in a single interaction (this is discussed in detail in [5–8] including a discussion of the corresponding information measures: pace, friction, and quality);
- broad fitness: the ability to achieve favorable outcomes over multiple interactions, potentially of different types;
- adaptiveness: the ability to achieve favorable outcomes when the nature of interactions available in the environment changes.

Broad fitness takes into account factors that depend on multiple interactions. For example, there are many examples of machine learning in which human biases become evident over time [17,18]. These provide examples in which broad fitness can include ethical or social factors not always taken into account in narrow fitness or not evident in small numbers of interactions.

The degree of fitness depends on the component pattern of an IE. Here we are drawing on terminology used in IT architecture [12]. A component is a separable element of the IE—something that processes information in a particular way. In this sense, different applications and IT infrastructure are components for an organization; components for people are described in [19] (the authors say "inference, and cognition more generally, are achieved by a coalition of relatively autonomous modules that have evolved [. . .] to solve problems and exploit opportunities" and a "relatively autonomous module" corresponds to a component).

Figure 1 shows how these elements relate. In the figure, the superscripts 1, 2 and 3 refer to narrow fitness, broad fitness, and adaptiveness, respectively.

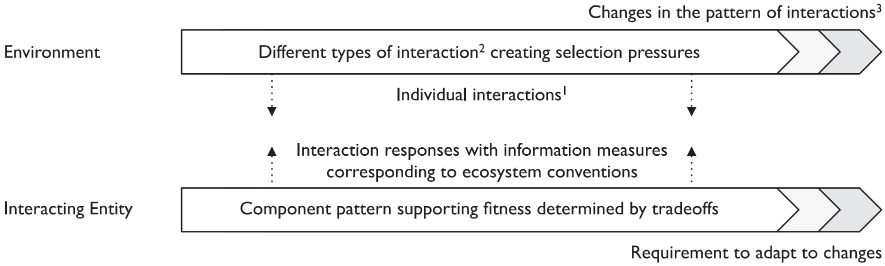

Changes in the pattern of interactions[3]
Environment
Different types of interaction[2] creating selection pressures
Individual interactions[1]
Interaction responses with information measures corresponding to ecosystem conventions
Interacting Entity
Component pattern supporting fitness determined by tradeoffs
Requirement to adapt to changes

Figure 1. Levels of interaction and fitness.

Selection pressures lead to the formation of information ecosystems [5]. Examples include English speakers, computer systems that exchange specific types of banking information, mathematicians, finance specialists and many others. Each ecosystem has its own conventions for exchanging and processing information. Within different ecosystems, modelling tools (using the term from [5–8]) such as languages, mathematics and computer protocols have evolved to structure and manipulate information within the ecosystem. An IE outside the ecosystem may not be able to interpret the information—think of a classical languages scholar trying to understand quantum mechanics.

Information relates to the physical world. Call a slice a contiguous subset of space-time. A slice can correspond to an entity at a point in time (or more properly within a very short interval of time), a fixed piece of space over a fixed period of time or, much more generally, an event that moves through space and time. This definition allows great flexibility in discussing information. For example, slices are sufficiently general to support a common discussion of nouns, adjectives, and verbs, the past and the future.

Slices corresponding to ecosystem conventions for representing information we call content with respect to the ecosystem. Content is structured in terms of chunks and assertions. A chunk specifies a constraint on sets of slices (e.g., "John", "lives in Rome", "four-coloring"). An assertion hypothesizes a relationship between constraints (e.g., "John lives in Rome"). Within ecosystems and IEs, pieces of information are connected in an associative model (for example, Quine's "field of force whose boundary conditions are experience" [20], the World Wide Web, or Kahneman's "associative memory" [4]) with the nature of the connections determined by ecosystem conventions.

The effect of competition and selection pressures over time is to improve the ability of IEs and ecosystems to process information corresponding to different measures of information [5]. The quality of information may improve, in the sense that it is better able to support the achievement of favorable outcomes; it may be produced with lower friction [21] or it may be produced faster. Or there may be more general tradeoffs in which the balance between quality, friction and pace varies.

Selection pressures ensure that information is generally reliable enough for the purposes of the ecosystem within the envelope in which the selection pressures apply. However, quality issues and the limitations discussed below mean that outside this envelope we should not expect ecosystem conventions to deliver reliable results [6–8]. This is particularly important in an era of rapid change, such as the current digital revolution, in which IEs cannot keep pace with the change—for example creating the "digital divide" for people [22,23] and less market success for businesses [11]. For people, ecosystems can be age-related—for example, "digital natives", "digital immigrants" and "digital foreigners" [24] differ in their approach to the use of digital information.

2.1. AI and Machine Learning

AI is causing much debate at the moment. On the one hand it promises to revolutionize business [25] and on the other it may help to trigger the singularity [1,2,26]. The major recent developments in AI have been in machine learning—Domingos provides an overview in [27].

In this paper, we are concerned with the relationship between AI and information (as described in the previous section). As [25] demonstrates, AI can impact many elements of information processing for organizations. Importantly, it can make a significant improvement to all levels of fitness but to turn this into benefits, an implementation for an organization needs to link a detailed understanding of the three levels of fitness, their relationship and how each AI opportunity can improve them. In turn, this requires an understanding of measures of information such as friction, pace, and quality [5,6]. These points are expanded below.

2.2. Capability Requirements

To help understand how IEs can provide the levels of fitness required to thrive we can draw a capability model using a technique from enterprise architecture [12]. This approach is an elaboration of the approach taken in [5–8]. A capability is the ability to do something and we can draw a capability model for information capabilities, as in Figure 2, using the three levels of fitness identified in the previous section. Please note that this is a generic capability model that applies to all IEs and the degree to which capabilities are present in any IE may vary hugely. There are many other such models (for example, Figure 5 in [15]) highlighting different viewpoints but Figure 2 focuses on the issues that relate to fitness.

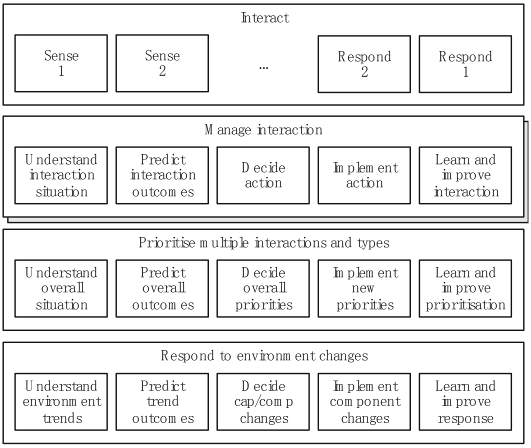

Interact
Sense 1
Sense 2
...
Respond 2
Respond 1
Manage interaction
Understand interaction situation
Predict interaction outcomes
Decide action
Implement action
Learn and improve interaction
Prioritise multiple interactions and types
Understand overall situation
Predict overall outcomes
Decide overall priorities
Implement new priorities
Learn and improve prioritisation
Respond to environment changes
Understand environment trends
Predict trend outcomes
Decide cap/comp changes
Implement component changes
Learn and improve response

Figure 2. Information capability model.

An IE needs the capability to interact and, in turn, this needs the ability to sense and respond to the environment (for example, to understand speech and to talk). To manage the different levels of fitness it needs to be able to:

- manage each interaction and to decide how to respond (if at all)—each type of interaction may require different specific capabilities;
- prioritize the response to different interactions of the same or different types;
- respond to environment change—this is a priority for businesses as the world becomes increasingly digital and the implementation of AI and machine learning takes hold [11].

In each case there is a five steps process that applies at the appropriate level, involving:

- understand the situation—identifying what is relevant (distinguishing signal from noise), interpreting the relevant information in the environment by connecting it to information in memory, and distilling it to an appropriate level for analysis and decision-making;
- link the situation to potential outcomes and understand their relative favorability;

- decide how to respond;
- implement the change—in information terms this means converting the decision into instructional information;
- learn and improve.

Each capability describes what an IE could do but not how it does it or the degree to which it does it. Any particular IE will have a set of components conforming to a component pattern that provides the capabilities. The nature of component patterns is discussed in Section 3.

3. The Limitations of Information and Applications of AI to Business

Information and its processing is subject to many limitations—these are discussed at length in [5–8]. These limitations occur because it is difficult linking environment states with future outcomes and the required actions to achieve favorable outcomes under the influence of selection pressures. The impact is always that perfection is impossible and under different selection pressures there are tradeoffs with respect to the different components of fitness.

This section provides an overview of these problems and how different ecosystem conventions and modelling tools can help to overcome them. In particular, it discusses the impact on the problems of AI and how the use of AI can, in some cases, help to resolve them.

The first problem is combinatorial. The number of possible environment states, outcomes, and the relationships between them is huge and in each interaction, these must be boiled down by an IE to a single action (including the possibility of inaction). The basis for overcoming this problem is provided fundamental characteristics of information—symbols and the means of associating them in different ways.

The second problem is how to make the tradeoffs between the information measures (like pace, friction, and quality [5,6]) required to support favorable outcomes. This problem breaks down into several sub-problems:

- what properties to measure in the environment and to what quality?
- how to overcome the complexities involved in interpretation, inference, and instruction—how to develop shortcuts in the provision of the required quality?
- how to overcome contention at different levels—between different information measures, between the needs of the present and the needs of the future, between the needs of different interaction types, between different ecosystems (especially at ecosystem boundaries) and how to keep the IE aligned with fitness even as the requirements of the environment change?
- how to challenge the level of quality achieved—how the ecosystem can apply selection pressures of its own to ensure quality?

The final problem is architectural (in the sense of [12])—what component pattern is best and how should this change over time?

The ways in which these problems are resolved in different ecosystems determine the ecosystem conventions and the detailed selection mechanisms that apply.

3.1. The Combinatorial Problem

Information overload has been much discussed [28] but this is but one symptom of a deeper problem: there is an unimaginably large number of measurable potential environment states, potential outcomes, and connections between them.

An environment state or, indeed, any slice, even if measured with relatively poor quality, is not easy to manipulate and process—there are a (potentially large) number of properties and their values to consider. Therefore, there is a large processing saving (and reduction in friction and pace) if a simple identifier, associated with the slice, is used instead. If the identifier is connected, in some way, with the slice and it is clear what the identifier means (by reference to the slice properties, as needed), then processing will be simplified hugely. Therefore, it is unsurprising that identifiers are widespread in information storage and processing in the form of symbols or sets of symbols. The nature of the symbol is not relevant (and the fact that symbols can be arbitrary is a fundamental principle of semiotics [29]). What matters is that the symbol can be connected, as needed, to the slice it is connected to and that it can be discriminated from other symbols. (Please note that the requirement that symbols can easily be discriminated foreshadows one of the benefits of the digital world—see the discussion in [30].)

This helps to solve the processing problem but if we need a symbol for each possible slice then we have not escaped the combinatorial problem entirely. It would also be useful for a symbol to apply to a set of slices that have something in common—that meet some set of constraints. This is, for example, the way language works: verbs relate to sets of event slices with common properties; adjectives relate to sets of slices with some common properties and so forth.

Set inclusion is binary: in or out. Therefore, by taking this route to solve the combinatorial problem, the use of symbols has built in a fundamental issue with information that underpins many of the limitations analyzed in [5–8]. The authors discuss this question in their analysis of patterns in [28] and say: "It is paradoxical that the similarity of the elements of a set creates a difference between the very elements of the set and all of the things not in the set". If one or more pieces of content maps close to the boundary of a set (in the sense that a small change in property values moves it to the other side of a boundary) then an interpretation, inference or instruction that relies on that positioning requires the quality to be high enough to guarantee the positioning. Call this the discrimination problem. An extreme form of the discrimination problem arises from chaotic effects [31] in which arbitrarily small changes can give rise to large outcomes. As demonstrated in [5,6], much routine information processing ignores this question entirely. In machine learning terms, the discrimination problem translates into the levels of risk and tolerance associated with false positives and false negatives [32].

The use of symbols enables another trick: symbols can relate to other symbols not just to sets of slices (this is because symbols correspond to sets of slices conforming to constraints in a particular ecosystem [5]). Therefore, as described in [6,7], we need to be careful to distinguish between content slices—those interpreted as symbols in an ecosystem (by IEs in the ecosystem)—and event slices—those that do not.

Of course, all the discussion about symbols is ecosystem-specific. A symbol in one ecosystem may not be one in another—words in one language may not be in another language, mathematics is meaningless to non-mathematicians.

The combinatorial challenge is magnified when we consider multiple interaction types and environment change. Multiple interaction types may need more slice properties, more symbols and, perhaps, different ecosystems and ecosystem conventions. In addition, recognizing environment change requires the ability to store and process historical data that will allow the identification of trends (access to this historical "big data" has been one of the drivers of machine learning).

This leads to another aspect of the combinatorial challenge: how should information and components be structured to enable fitness at the various levels (including adaptiveness). Remembering that information is about connecting states, outcomes and actions, there is a key structuring principle here (used commonly in the technology industry [9]). Decoupling two components enables one to be changed without changing the other (decoupling is discussed more in the discussion about component patterns below) and this requires them to be separable in some sense. We can replay the discussion above in the following way:

- The use of symbols separates information from source slices and their properties;
- Ecosystem conventions separate symbols from particular slice representations (so words can be written or spoken, for example);
- Evolving ecosystem conventions separate processing (and the making of connections) from particular IEs (so computers can automate some human activities, for example);
- Communication separates content from a physical location (so content can be duplicated at distance).

In this way, the evolution of ecosystem conventions progressively frees up information from the particular process that generates it. This progression is neatly reflected in the development of organizational enterprise architectures [12] in which two major themes have emerged:

- Developments in data warehouses, business intelligence, data engineering, data lakes and data hubs enable the collation and manipulation of data from many different sources;
- Digital technologies enable information to be available at widely different times and places and on many different devices.

One strategy for addressing the combinatorial problem is increasing processing power and this is precisely what Moore's law [33] has provided for machine learning (combined with access to access to large volumes of data—so-called ”big data”). This increase in power and access to data has been one of the drivers of the current boom in AI but is, as yet, a considerable distance away from resolving the combinatorial problem, even aside from the other difficulties outlined below.

3.2. Selection Tradeoffs, Viewpoints and Rules

The impact of the combinatorial problem is that information processing uses a strict subset of the properties of environment states available, makes quality tradeoffs and may be linked to a strict subset of possible outcomes. In other words, all information processing has a viewpoint (using the terminology employed in [7,8]). This is routine in day-to-day life—for example:

- with the same evidence, different political parties reach very different conclusions about the right course of action in any case;
- in legal cases, the prosecution and defense represent different viewpoints;
- even in science, there are divisive debates about the merit of particular hypotheses (this is represented, for example, in Kuhn's philosophy of science [34]).

Since these viewpoints are inevitable, we need to understand their impact. This is the focus of the following sections.

3.2.1. Measurement

Measurement is about converting environment states into properties and values or more abstract content (subject, of course, to the prevailing ecosystem conventions). How does this relate to fitness measures?

One dimension is the number of properties measured, how they are measured and the quality of the measurement. In addition, once properties are measured, how often do they need to be re-measured—to what extent is timeliness an issue [5]?

When multiple types of interaction are considered, an extra dimension comes into play—to what extent can measurement required for one interaction type be used for another—if the different interactions use different ecosystem conventions, can the properties be measured and processed in the same way and what are the implications if they are not? This a common problem in organizations—the quality of information needed to complete a process successfully may be far less than that required for accurate reporting.

Finally, when the environment is changing, there may be a requirement for new properties to be measured or for changed ecosystem conventions to be considered.

Machine learning can be one of the drivers behind improved measurement for organizations because the recognition of patterns and its automation [27] are fundamental principles in the discipline [32]. Machine learning can improve pace, reduce friction and, in some cases, improve quality also through the automation of learning based on good quality data (although there have been some significant difficulties [17,18]).

3.2.2. Information Processing Limitations and Rules

As discussed in [5–8], different strategies are possible for information processing depending on the degree to which each of quality, pace or friction is prioritized in terms of narrow fitness. A rigorous process focusing on quality requires an approach such as that of science but many ecosystems cannot afford this overhead. Instead they rely on rules that exploit the regularities in the environment, as discussed by the authors in [19], who say:

> *"What makes relevant inferences possible [. . .] is the existence in the world of dependable regularities. Some, like the laws of physics, are quite general. Others, like the bell-food regularity in Pavlov's lab, are quite transient and local. [. . .] No regularities, no inference. No inference, no action."*

There can be difficulties associated with exploiting these regularities both for people and machines. As Kahneman points out [4] with respect to our innate, subconscious responses (what he calls System 1): "System 1 is radically insensitive to both the quality and quantity of information that gives rise to impressions and intuitions." As Duffy says in [35] "and the more common a problem is, the more likely we are to accept it as the norm".

Machine learning [27] finds and exploits some of these regularities but has been subject to some well-publicized issues associated with bias [17,18] (although the biases revealed have, in some cases, been less than people display [18]).

The nature of the regularities is discussed in [8] in which inference is categorized in terms of:

- Content inference—using only the rules associated with a particular modelling tool (for example, formal logic or mathematics);
- Causation—in which inference is based on one or more causation processes;
- Similarity—in which inference is based on the similarity between sets of slices and the assumption that the similarity will extend.

Machine learning is based on similarity, so this categorization poses a question. For what types of information processing is machine learning the most appropriate technique and when are other techniques appropriate? In particular, when is simulation (concerned with modelling causation) more appropriate? This question is discussed in Section 4.

Content processing has clear benefits in terms of friction and pace—making the connection with events incurs much higher friction (this is the relationship between theoretical physics and experimental physics, for example, and consider the cost of the Large Hadron Collider). Wittgenstein also referred to this idea and the relationship between content and events [36,37] with respect to mathematics:

"[I]t is essential to mathematics that its signs are also employed in mufti";

"[I]t is the use outside mathematics, and so the meaning ['Bedeutung'] of the signs, that makes the sign-game into mathematics".

An equally insidious shortcut is output collapse (to use the term used in [8]). There are uncertainties about interpretation, inference and instruction caused by information quality limitations. However, an interaction results in a single action by an IE (where this includes the possibility of no action at all) and examining a range of potential outcomes and actions increases friction. Therefore, in many cases, interpretation and inference are designed to produce a single answer and the potentially complex distribution of possibilities collapses to a single output. If this collapse occurs at the end of the processing, then it may not prejudice quality. However, if it occurs at several stages during the processing then it is likely to.

There is another type of shortcut. This is quality by proxy in which quality is assessed according to the source of the information (linked to authority, brand, reputation, conformance to a data model or other characteristics). In [38], the authors express this idea elegantly with respect to documents: "For information has trouble, as we all do, testifying on its own behalf... Piling up information from the same source does not increase reliability. In general, people look beyond information to triangulate reliability."

As a result, of selection tradeoffs, these various types of shortcut become embodied in processing rules that are intended to simplify processing with sufficient levels of quality. The rules are defined with a degree of rigor consistent with ecosystem conventions (for example, rigorous for computer systems but less so for social interaction).

Organizations use rules such as this (called business rules) routinely. Business processes embody these business rules in two senses. At a large scale, a process defines the rules by which a business intends to carry out an activity (for example, how to manage an insurance claim). In addition, in a more detailed sense, business rules capture how to accomplish particular steps (for example, the questions to ask about the nature of the claim). Machine learning can improve both of these aspects. In the first case, the context of the process (for example, information about the claim) may change the appropriate next step (for example, the appropriate level of risk assessment to apply). Therefore, rather than a fixed set of steps as captured in a process map, the process may become a mixture of fixed steps and something akin to a state machine [39] or, in some cases, just a state machine. This change relies on a continuous situation awareness (as described in Figure 2) that can use machine learning as a measurement tool. In addition, machine learning can also refine the business rules over time based on the developing relationship between the rules and fitness objectives (for example, the tradeoff between quality and friction or pace). It may be appropriate to change the rules (changing the questions to ask in this example) when more information is learnt about the effectiveness of the rules or it becomes possible to tune the rules more specifically to individual examples.

3.2.3. Contention

Selection tradeoffs are about managing contention and ecosystem conventions embed the tradeoffs. For a single interaction there is contention between pace, friction, and quality. This type of contention is discussed in detail in [5–8].

Multiple interactions and types of interaction introduce extra dimensions. The first is between the present and the future: how much should an IE optimize the chances of a favorable outcome for a single interaction against the possibilities of favorable interactions in the future? The second is between different interaction types: how much should an IE focus on one type of interaction compared to others? Or, to put it another way, how much should the IE specialize? Many authors in different disciplines have discussed specialization as a natural outcome of selection pressures—for example:

- Philosophers from Plato [40] onwards discussing the division of labor;
- Biologists including Darwin [10], since species themselves are examples of specialization;
- Business writers discussing differentiation, including Porter [41].

More generally, there might be what we can call conflict of interest between narrow fitness and broad fitness especially when the nature of quality associated with narrow fitness does not match that associated with broad fitness. In [42], the author gives examples of the impact of conflict of interest on science. There have been several well-publicized examples concerning machine learning [18]. In these cases, narrow fitness is defined in terms of the data used to generate the learnt behavior but the data itself may embed human biases. As a result, narrow fitness (linked to training data) does not take ethical and social issues into account and broad fitness is reduced.

The next point of contention arises from ecosystem boundaries. The conventions that apply on one side of the boundary may be very different from the other (we only need to consider speakers of different languages or the user experience associated with poorly defined web sites) and there may be contention at fundamental levels. One initial driver of AI (the Turing Test [43]) was aimed at testing the human/computer ecosystem boundary. This is still of considerable importance but a related question in organizations is understanding how AI and people can work together [44] and how AI can support other ecosystem boundaries.

Finally, there may be contention in the balance of the selection pressures as the environment changes. For example, in the digital revolution engulfing the world of business [11] the balance between friction, pace and quality is changing—the ability to respond fast (i.e., pace) is becoming more important. Machine learning plays a part here since it is a mechanism for constantly re-learning from the environment.

3.2.4. Challenge and Assurance

For an IE, information processing is reliable if it helps to achieve a favorable-enough outcome—if the IE can rely on the processing within the envelope provided the ecosystem selection pressures (as discussed in [6–8], outside this envelope is it not guaranteed to be reliable enough). Therefore, how can ecosystems apply their own selection pressures to improve the reliability of information processing? An element that many ecosystems have in common is that of challenge. Table 1 (copied from [8]) shows some examples.

Table 1. Challenge.

Ecosystem	Hypothesis	Challenge
English criminal law (prosecution)	The defendant is guilty	The defense (plus, potentially, the appeals process)
Science	A prediction made by a hypothesis is true	Experiments to refute or confirm the prediction
Mathematics	A theorem is proved	Peer review
Computer systems	The system will perform as required	Tests that the system meets its requirements

The objective of each challenge is to identify weaknesses in information processing either in terms of its output (e.g., refutation in scientific experiments), the input assertions on which it is based (e.g., the evidence in a trial) or the steps of the inference (e.g., peer review in mathematics).

The generic mechanism is similar in each case. A related ecosystem has selection pressures in which favorable outcomes correspond to successful challenges. The degree to which the challenge is rigorous depends on the selection pressures that apply to it and, in some cases, the degree to which a different IE from the one making the inference conducts it (to avoid the conflicts of interest discussed in [42], for example).

Therefore, given that challenge is a type of selection pressure, how does the nature of challenge relate to fitness criteria? There are some obvious questions. First, is the inference transparent enough to be amenable to challenge? This is one of the questions that has been raised about deep learning although recent research has started to address this question [45].

Secondly, what is the degree of challenge—how thorough is it? This is an important issue addressed by organizations as they implement machine learning—how does the assurance of machine learning relate to conventional testing and are additional organizational functions required. This is discussed below.

Thirdly, what is the scope of the challenge is relation to fitness—is it concerned with narrow fitness or does it incorporate broad fitness and adaptiveness as well? This is one of the considerations described in detail with respect to technology in [9]; but the issue as applied to machine learning is more extensive because machine learning learns from historic data that may not encapsulate the desired requirements of broad fitness and is unlikely to include the requirements of adaptiveness.

Challenge and assurance is important for machine learning since there are many public examples in which machine learning has delivered unacceptable results [17,18]. An element of broad fitness that has been the subject of much attention is ethics [46], because of these issues and also the long-term direction of AI and the potential singularity [1,2,26].

The purpose of the challenge is to identify what the software industry calls test cases [39]—a set of inputs and outputs designed to cover the range of possibilities thoroughly enough to provide confidence of reliability (in the context of the ecosystem conventions). In clearly defined domains such as Go and chess, the test cases themselves can be generated by machine learning but where there is a level of organizational risk involved (e.g., reputational, ethical, operational or security-related) then more traditional forms of assurance may be required focusing on the training data, the selection of a range of scenarios to test and an organizational assurance function to analyze examples of the discrimination problem and potential impacts. Since machine learning can re-learn periodically, these forms of assurance may need to be applied, in some form, regularly.

Therefore, we can conclude that, as AI becomes more prevalent and the issues discussed above become more important, organizations will need to understand and manage the potential impacts and risks. This will require an organizational assurance function that will ensure that the right degree of challenge is applied and analyze and, where necessary, forecast the impact of AI on business results.

3.3. Component Pattern

Components are the physical realization of capabilities (see Figure 2) and components can be arranged in different patterns. Table 2 shows some examples of components. The relationship between capabilities and components for business and technology architectures is part of the day-to-day practice for enterprise architects [12]. The development of component patterns to meet future fitness requirements is a key part of developing future architectures to support organizational fitness requirements [9]. We can use these ideas to analyze component patterns for IEs.

Table 2. Examples of components.

IE	Interaction Component	Interpretation, Inference, and Instruction Component
People	Senses (eyes, . . .)	Different brain mechanisms (see [19])
Organizations	Sales people, customer research, web sites, . . .	Different organizational functions and their supporting computer systems (for example, qualifying sales opportunities, deciding the chances of winning and deciding how to price the product or service)
Computer system architectures	Virtual assistants (e.g., Alexa, Siri), apps, enterprise applications, security intrusion detection, . . .	Algorithms, machine learning tools

Components evolve incrementally and become integrated to meet the need to connect environment states to outcomes and actions. The nature of the integration and the pre-dominance of certain components can imply different patterns. These patterns have a set of characteristics based on the capabilities shown in Figure 2:

- Channel-aligned: in this case, interaction components extend to encompass wider information processing. For example, Pinker [47] gives many examples in which the processing of the human brain is influenced (and constrained) by language and, indeed, some believe that language processing underpins all of human thought (for example in [48] the author says "I believe that language is also the medium by which we formulate our conceptual thinking. I regard thinking as silent language.").
- Function-aligned: in this case, components (like interaction components) are built out from particular functions. For example, many organizational capabilities (such as finance and HR) are supported by software products that have developed from a functional base and also provide interaction (e.g., through web sites) and analytics.
- Multi-function: in this case, different components providing different functions are integrated together. For example, in [19], the authors make the case that many specialized inference mechanisms have evolved in people; they say: "inference, and cognition more generally, are achieved by a coalition of relatively autonomous modules that have evolved [. . .] to solve problems and exploit opportunities". Another example is an extension of the previous case in which organizations have specific components to support finance, HR, manufacturing, retailing and other organizational functions.

- Information-aligned: in this case, components are based on the capability model in Figure 3. For example, many organizations have built data warehouses to support business intelligence as well as data lakes and analytics capabilities [49].

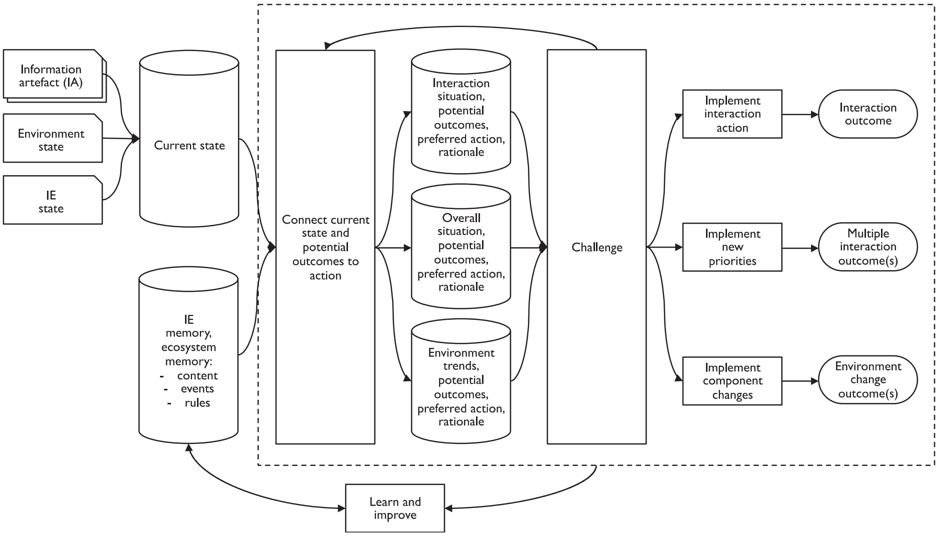

Information artefact (IA)
Environment state
IE state
Current state
IE memory, ecosystem memory:
- content
- events
- rules
Connect current state and potential outcomes to action
Interaction situation, potential outcomes, preferred action, rationale
Overall situation, potential outcomes, preferred action, rationale
Environment trends, potential outcomes, preferred action, rationale
Challenge
Implement interaction action
Implement new priorities
Implement component changes
Interaction outcome
Multiple interaction outcome(s)
Environment change outcome(s)
Learn and improve

Figure 3. Information-aligned component pattern.

These different types of pattern have different strengths and weaknesses based on the core components. IEs often have a combination of these patterns and the balance between them impacts elements of fitness. Component patterns embed the information structure and processing tradeoffs implicit in ecosystem conventions and these both enable and constrain different elements of fitness. For example, channel-aligned patterns are strong when interaction is a large element of fitness; information-aligned patterns are strong when information needs to be integrated separately outside the processes that generated the information.

However, component patterns may need to change. For example, there is a clear trend [30] for organizations to respond to the digital economy by adding an information-aligned pattern that takes advantage of machine learning. Figure 3 shows a generic information-aligned component pattern.

A more extreme example of change is the trend towards the AI-assisted human and the need for humans and AI to work together [44].

Components need to be integrated in order to link environment states to outcomes and actions. Narrow fitness demands short and efficient processing embedding rules that deliver sufficient quality. Broad fitness requires additional processing complexity and may also require the integration of different ecosystems with different conventions. Both of these are drivers for tight integration between components.

However, adaptiveness requires decoupling—the ability to change components independently [9]—because otherwise change incurs too much friction. This generates a tension between the different types of fitness; without a sufficiently strong adaptiveness selection pressure, the nature of the component integration can be brittle and resist change.

For organizations, machine learning has a role to play here. If some or all of the business rules are based on machine learning, then periodic re-learning can update the rules (but see the discussion about re-learning below). For this to be the case, the organization will need a component pattern that is sufficiently information-aligned. As AI becomes embedded in more and more technology, the shift towards information-alignment, or the addition of information-alignment, will become more and more important.

The same change (towards information-alignment) is also true of quality improvement. Better-informed people make better decisions and the same principle underpins the implementation of machine learning in business. Improvements in interpretation and inference quality require richer access to information [5,8] that channel-alignment or function-alignment alone cannot provide.

In [19], the authors demonstrate that human inference has many different inference patterns. In addition, the mind does a wonderful job of giving us the illusion that things are well integrated even when, underneath, they are not; this is what magic relies on [50]. Machine learning may be heading in the same direction—current developments in AI contain several different patterns. Domingo [27] categorizes these as symbolists, connectionists, evolutionaries, Bayesians, and analogizers. However, more generally, AI is becoming a set of techniques embedded in numerous applications using whatever technique(s) is appropriate in each case. In this case, the integration question takes on another dimension: how can the interpretations and inferences of multiple components including AI integrate into reliable interpretations and inferences for the organization as a whole. The different components may use different ecosystem conventions with different information structures and process tradeoffs. There may be gaps between their domains (as in the magic example above). Other problems identified above (output collapse and contention) may apply. There is also an uneasy relationship between AI component integration and the discrimination problem. If inference relating to a critical boundary condition relies on integration between machine learning components, then the reliability of the integration needs to be tested rigorously.

The challenge becomes greater when we consider re-learning. One of the advantages of machine learning is that rules can be re-learnt as the environment changes. However, when many machine learning components are integrated to support a complex set of business functions, how should this re-learning work? Again, the principle of decoupling applies—we want the different interpretations and inferences to be independent. However, how do we know that this is the case? With more data or a change in the environment, new patterns may emerge in the data (that, after all, is the whole point of re-learning) and these new patterns may create new dependencies between the rules. This reinforces the need for assurance (as discussed above).

Therefore, we can conclude that, as machine learning becomes more pervasive, integrating different approaches to machine learning, each supporting different viewpoints and ecosystem conventions, will provide challenges in the following four areas:

- Providing high quality, coherent descriptive, predictive, and prescriptive information from disparate components each learning in different ways from different subsets of data at different times;
- Tackling the discrimination problem especially where components need to be integrated;
- Ensuring that content information processing does not suffer from the same limitations as for humans;
- Ensuring that the underlying data is of the required quality for each component.

These challenges provide a foretaste of the deeper issues with AGI discussed in the next section.

4. The Limitations of Information and AGI

AGI is one of the main factors driving AI research (see, for example, [26]) and, in the view of many authors (for example, [1,2]), AGI is a step on the road to the singularity. Therefore, it is important to understand the impact of the limitations of information and the theoretical and practical difficulties that they imply about AGI.

In this section, we discuss the following challenges for AGI based on the analysis above:

- How is fitness for AGI determined?

- How will AGI handle the integration of components, the need to accommodate different ecosystem conventions and be sufficiently adaptive?
- How will AGI process and relate abstractions and will it be able to avoid the difficulties that humans have with the relationship between abstractions and information quality?

One difference between narrow AI and AGI is that AGI needs to handle many interaction types and combinations of them, so how is it possible to define or characterize all of them? And how can we apply the right selection pressures—to use the terminology of IT, how can we define all of the test cases required? One approach of the AI community is to use AI techniques (like adversarial generative networks) to this further question. However, for difficult questions, and for broad fitness in general, at some stage people will need to be sure of the potential outcomes, so people will need to apply the right selection criteria even to those further AI techniques. It is difficult to see this as other than another manifestation of the combinatorial problem but magnified by the number of different interactions types and their combinations. Defining broad fitness for people and organizations includes the legislation of a country as well as cultural and moral imperatives, so how can we define it for AGI? (This topic has been recognized widely including by such multi-national bodies as the World Economic Forum who ask the question "How do we build an ethical framework for the Fourth Industrial Revolution" [51]). As well as these aspects, broad fitness for AGI will require rigorous security fitness. The combination of all of these is a dauntingly large task.

This implies that it is very difficult to define even what AGI is in enough detail to be useful in practice. In addition, we need a specific definition because overcoming the discrimination problem requires appropriately high information quality—for AGI, the discrimination required may include many issues concerning human safety, as we have already seen with autonomous cars.

One way round this is the AGI equivalent of "learning on the job"—allowing AGI to make mistakes and learn from them in the real world. Whether or not this is feasible depends on the fitness criteria that apply—it is difficult to see that this would be acceptable for activities with significant levels of risk. It has already caused reputational damage in the case of simple, narrow AI [17,18]. In [52], the authors address this issue when they ask the question: "why not give AGI human experience"? They then show how human experience is difficult to achieve. Given the discussion in Section 3 about viewpoints, if the experience of AGI is different from human experience then, necessarily, its viewpoint will be different and its behavior will be correspondingly different.

How about integration? In humans, different types of interpretation and inference use different components [19]. Currently, the same is true of machine learning—increasingly, it is a computing technique that is applied as needed. Therefore, it seems likely that AGI will need to integrate many different learning components. Domingos [27] suggests one integration approach and there are other approaches (e.g., NARS [53,54]). There are several issues here: ecosystem conventions, content inference, selection tradeoffs and component patterns.

Just as people may engage with different ecosystems (e.g., different languages, different organizational functions, computer systems, different fields of human endeavor (sciences, humanities)) AGI will need to be able to deal with different ecosystems and their relationships. Different ecosystems have different conventions and fitness criteria so AGI will need to manage these and convert between them. Again, the discrimination problem raises its head—different ecosystem conventions are not semantically interoperable. Combining processing using different ecosystem conventions risks what [7,8] refer to as "interpretation tangling" or "inference tangling" in which conventions that apply to one ecosystem (e.g., mathematics) are implicitly assumed to apply to another (e.g., language) resulting in unreliable results. A learning approach could only address these issues if the combinatorial problem described above does not apply (and in reality, it may not be possible even to identify or source all the possible combinations to learn).

Deep learning uses layers of neural networks in which intermediate layers establish some intermediate property and subsequent layers use these abstractions; thus, these subsequent layers are then using content processing. Metalearning [27] provides another example of content processing.

In these examples, because they apply to narrow AI, the limitations of content processing described in Section 3 have little impact. However, when we scale up to AGI with many components of different types developed for different ecosystems providing abstractions that are integrated by one or more higher levels of machine learning then the limitations of content processing may become a problem.

Content processing is used by ecosystems because the use of content rules is much faster and more efficient than event processing (testing against the properties and values of sets of slices)—this is an outcome of the combinatorial problem. Therefore, is it feasible that this requirement not be present for AGI? Only if the AGI could relate all information processing to events (not content) as it was needed. In the face of the discrimination problem this amounts to the ability to provide processing power to overcome much of the combinatorial problem. Even if Moore's law [33] continues, this is a difficult proposition to accept for the foreseeable future and even if it was feasible, there is no guarantee that it would not be subject to selection tradeoffs.

Therefore, we can conclude that content processing will likely be a part of AGI and therefore that the limitations of content processing will also apply and that, as a result, information quality will be compromised. However, without a definite AGI model to base the analysis on, the impact of this is unclear.

What about adaptiveness? Adaptiveness is, partly at least, an attribute of the component pattern. However, the experience from the technology industry, most recently in developing digital enterprise architectures [55] is that developing new component patterns is a change of kind not of degree—component pattern changes are difficult to evolve by small degrees. Thus, we cannot expect linear progress. This is discussed in [52] in which the authors include the following quote from [56] "The learning of meta-level knowledge and skills cannot be properly handled by the existing machine learning techniques, which are designed for object-level tasks". Perhaps AGI will need the ability to learn about component patterns themselves—when a new component pattern is needed the AGI will need to recognize it and evolve a new one; but even if this is feasible, where will the data come from?

In principle, AGI could be adaptive, within the context of a single component pattern, because it can re-learn periodically. However, re-learning will be subject to selection pressures and the possibility of tradeoffs and different ecosystem conventions. Thus, in practice, different machine learning components may re-learn at different rates and times raising the possibility of inaccuracies and inconsistencies exacerbating the discrimination problem and quality in general.

As Section 3 points out, the degree of decoupling within the component pattern is important for adaptiveness. The human brain masks the cognitive integration difficulties we all have [50] between different components. It is possible that this type of integration difficulty is a natural consequence of the tradeoffs between adaptiveness and other levels of fitness. Can we be sure that the same does not apply to AGI?

The discussion about information processing in Section 3 (and [8]) highlights another potential difficulty with machine learning and AGI. One of the prevalent ideas in technology at the moment, driven partly by the Internet of Things and the ability to understand the status of entities, is that of the "digital twin"—a simulation of those entities. Similar ideas are driving technologies such as virtual reality and, of course, in many scientific and other fields, simulation has long been a critical tool. Bringing these ideas together will support the creation of models of the environment enabling a richer simulation of external activities, leading to the question: under what circumstances will simulation be preferable to AI and how can they work together?

Machine learning exploits "the existence in the world of dependable regularities". However, will these dependable regularities occur reliably enough in the information available to machine learning to provide sufficient quality to overcome the discrimination problem? Might not inference based on causation be required to address some difficult instances of the discrimination problem? This question is the AI equivalent of the "blank slate" issue discussed by Chomsky [57] and many others. Since complex simulation relies on complex theoretical models, inference based on causation it is not, in the foreseeable future, amenable to machine learning.

5. Conclusions

The analysis of fitness and the limitations of information above provide a sound theoretical basis for analyzing AI both for implementation in organizations now and with respect to AGI. This analysis is validated by the current experience of AI and can also be used to define the following important implementation principles.

- Fitness: AI can make a significant improvement to all levels of fitness but to turn this into benefits the implementation of AI for organizations should be based on a detailed understanding of the three levels of fitness, the relationship of the levels and how each AI opportunity can improve them. In turn, this requires an understanding of measures of information such as friction, pace, and quality.
- Integration: Organizations will need to analyze the integration challenges of different AI approaches. As AI becomes more pervasive, integration will provide challenges in the following four areas:
  - ○ Providing high quality, coherent descriptive, predictive, and prescriptive information from disparate components each learning from different subsets of data at different times using different techniques;
  - ○ Tackling the discrimination problem especially where components need to be integrated;
  - ○ Ensuring that content processing does not suffer from the same limitations that it has for humans;
  - ○ Ensuring that the underlying data is of the required quality for each component.
- Ecosystem boundaries: One initial driver of AI (the Turing Test) was aimed at the human/computer ecosystem boundary. This is still important but a related question in business is understanding how AI and people can work together and how AI can support other ecosystem boundaries.
- Assurance: As AI becomes more prevalent and the issues discussed above become more important, organizations will need to understand and manage the potential impacts and risks. This requires an organizational assurance function that will analyze and, where necessary, forecast the impact of AI on business results.

These topics increase in importance with respect to AGI because the theoretical difficulties will become more profound. The following questions highlight important theoretical difficulties for which AGI research will require good answers:

- How is fitness for AGI determined?
- How will AGI handle the integration of components, the need to accommodate different ecosystem conventions and be sufficiently adaptive?
- How will AGI process and relate abstractions and will it be able to avoid the difficulties that humans have with the relationship between abstractions and information quality?

When we analyze these questions, it is clear that there are difficult information theoretic problems to be overcome on the route to the successful implementation of AGI.

Funding: This research received no external funding.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Bostrom, N. *Superintelligence: Paths, Dangers, Strategies*; Oxford University Press: New York, NY, USA, 2014.
2. Tegmark, M. *Life 3.0: Being Human in the Age of Artificial Intelligence*; Knopf Publishing Group: New York, NY, USA, 2017.

3. Bringsjord, S.; Govindarajulu, N.S. Artificial Intelligence. *The Stanford Encyclopedia of Philosophy*. Zalta, E.N., Ed.; 2018. Forthcoming. Available online: https://plato.stanford.edu/archives/fall2018/entries/artificial-intelligence/ (accessed on 19 December 2018).
4. Kahneman, D. *Thinking, Fast and Slow*; Macmillan: London, UK, 2011.
5. Walton, P. A Model for Information. *Information* **2014**, *5*, 479–507. [CrossRef]
6. Walton, P. Measures of information. *Information* **2015**, *6*, 23–48. [CrossRef]
7. Walton, P. Information and Meaning. *Information* **2016**, *7*, 41. [CrossRef]
8. Walton, P. Information and Inference. *Information* **2017**, *8*, 61. [CrossRef]
9. Ford, N.; Parsons, R.; Kua, K. *Building Evolutionary Architectures: Support Constant Change*; O'Reilly Media: Sevvan, CA, USA, 2017.
10. Darwin, C. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*; John Murray: London, UK, 1859.
11. Westerman, G.; Bonnet, D.; McAfee, A. *Leading Digital: Turning Technology into Business Transformation*; Harvard Business Review Press: Cambridge, MA, USA, 2014.
12. The TOGAF Standard. Available online: https://publications.opengroup.org/standards/togaf (accessed on 19 December 2018).
13. Avgeriou, P.; Uwe, Z. Architectural patterns revisited: A pattern language. In Proceedings of the 10th European Conference on Pattern Languages of Programs (EuroPlop 2005), Bavaria, Germany, 6–10 July 2005.
14. Burgin, M. Principles of General Ecology. *Proceedings* **2017**, *1*, 148. [CrossRef]
15. Burgin, M.; Zhong, Y. Information Ecology in the Context of General Ecology. *Information* **2018**, *9*, 57. [CrossRef]
16. Floridi, L. *The Philosophy of Information*; Oxford University Press: Oxford, UK, 2011.
17. DeBrusk, C. The Risk of Machine-Learning Bias (And How to Prevent It). MITSloan Management Review. 2018. Available online: https://sloanreview.mit.edu/article/the-risk-of-machine-learning-bias-and-how-to-prevent-it/ (accessed on 19 December 2018).
18. Miller, A. Want Less-Biased Decision? Use Algorithms. Harvard Business Review. 2018. Available online: https://hbr.org/2018/07/want-less-biased-decisions-use-algorithms (accessed on 19 December 2018).
19. Mercier, H.; Sperber, D. *The Enigma of Reason*; Harvard University Press: Cambridge, MA, USA, 2017.
20. Quine, W.V.O. *"Two Dogmas of Empiricism", Reprinted in from a Logical Point of View*, 2nd ed.; Harvard University Press: Cambridge, MA, USA, 1951; pp. 20–46.
21. Gates, B.; Myhrvold, N.; Rinearson, P. *The Road Ahead*; Viking Penguin: New York, NY, USA, 1995.
22. Norris, P. *Digital Divide: Civic Engagement, Information Poverty and the Internet Worldwide*; Cambridge University Press: New York, NY, USA, 2001.
23. Government Digital Inclusion Strategy. 2014. Available online: https://www.gov.uk/government/publications/government-digital-inclusion-strategy/government-digital-inclusion-strategy (accessed on 19 December 2018).
24. Prensky, M. Digital Natives, Digital Immigrants Part 1. *On The Horizon* **2001**, *9*, 1–6. [CrossRef]
25. Manyika, J.; Bughin, J. The Promise and Challenge of the Age of Artificial Intelligence. McKinsey Global Institute Executive Briefing. 2018. Available online: https://www.mckinsey.com/featured-insights/artificial-intelligence/the-promise-and-challenge-of-the-age-of-artificial-intelligence?cid=eml-app (accessed on 19 December 2018).
26. Logan, R.K. (Ed.) Special Issue AI and the Singularity: A Fallacy or an Opportunity. Information. 2018. Available online: https://www.mdpi.com/journal/information/special_issues/AI%26Singularity (accessed on 19 December 2018).
27. Domingos, P. *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*; Penguin: London, UK, 2015.
28. Logan, R.K.; Tandoc, M. Thinking in Patterns and the Pattern of Human Thought as Contrasted with AI Data Processing. *Information* **2018**, *9*, 83. [CrossRef]
29. De Saussure, F. *Course in General Linguistics*; Bally, C., Sechehaye, A., Eds.; Duckworth: London, UK, 1983.
30. Walton, P. Digital information and value. *Information* **2015**, *6*, 733–749. [CrossRef]
31. Lorenz, E.N. Deterministic Nonperiodic Flow. *J. Atmos. Sci.* **1963**, *20*, 130–141. [CrossRef]
32. Sammut, C.; Webb, G.I. *Encyclopedia of Machine Learning*; Springer Science & Business Media: New York, NY, USA, 2011.

33. Moore, G.E. Cramming more components onto integrated circuits. *Electronics* **1965**, *38*, 114–117. [CrossRef]
34. Kuhn, T.S. *The Structure of Scientific Revolutions*, 2nd ed.; University of Chicago Press: Chicago, IL, USA, 1970.
35. Duffy, B. *The Perils of Perception: Why We're Wrong About Nearly Everything*; Atlantic Books: London, UK, 2018.
36. Wittgenstein, L. *Remarks on the Foundations of Mathematics*, Revised Edition; von Wright, G.H., Rhees, R., Anscombe, G.E.M., Eds.; Basil Blackwell: Oxford, UK, 1978.
37. Wittgenstein, L.; Bosanquet, R.G. *Wittgenstein's Lectures on the Foundations of Mathematics*; Diamond, C., Ed.; Cornell University Press: Ithaca, NY, USA, 1976.
38. Brown, J.S.; Duguid, P. *The Social Life of Information*; Harvard Business Press: Boston, MA, USA, 2000.
39. Sommerville, I. *Software Engineering*; Addison-Wesley: Harlow, UK, 2010.
40. Plato. *Plato's the Republic*; Books, Inc.: New York, NY, USA, 1943.
41. Porter, M.E. *Competitive Strategy: Techniques for Analyzing Industries and Competitors*; Free Press: New York, NY, USA, 1980.
42. Goldacre, B. *Bad Science*; Harper Perennial: London, UK, 2009.
43. Turing, A. Computing Machinery and Intelligence. *Mind* **1950**, *59*, 433–460. [CrossRef]
44. James Wilson, H.; Daugherty, P.R. Collaborative Intelligence: Humans and AI are Joining Forces. Harvard Business Review. 2018. Available online: https://hbr.org/2018/07/collaborative-intelligence-humans-and-ai-are-joining-forces (accessed on 19 December 2018).
45. Foy, K. Artificial Intelligence System Uses Transparent, Human-Like Reasoning to Solve Problems. MIT News. 2018. Available online: http://news.mit.edu/2018/mit-lincoln-laboratory-ai-system-solves-problems-through-human-reasoning-0911 (accessed on 19 December 2018).
46. Bostrom, N. *The Ethics of Artificial Intelligence (PDF)*; Cambridge University Press: Cambridge, MA, USA, 2011.
47. Pinker, S. *The Stuff of Thought*; Viking: New York, NY, USA, 2007.
48. Logan, R.K. Can Computers Become Conscious, an Essential Condition for the Singularity? *Information* **2017**, *8*, 161. [CrossRef]
49. Correia, R.C.M.; Spadon, G.; De Andrade Gomes, P.H.; Eler, D.M.; Garcia, R.E.; Olivete Junior, C. Hadoop Cluster Deployment: A Methodological Approach. *Information* **2018**, *9*, 131. [CrossRef]
50. Macknik, S.L.; Martinez-Conde, S. *Sleights of Mind*; Picador: Surrey, UK, 2011.
51. Sutcliffe, B.; Allgrove, A.-M. How Do We Build an Ethical Framework for the Fourth Industrial Revolution. Available online: https://www.weforum.org/agenda/2018/11/ethical-framework-fourth-industrial-revolution/ (accessed on 19 December 2018).
52. Wang, P.; Liu, K.; Dougherty, Q. Conceptions of Artificial Intelligence and Singularity. *Information* **2018**, *9*, 79. [CrossRef]
53. Wang, P. *Rigid Flexibility: The Logic of Intelligence*; Springer: Dordrecht, The Netherlands, 2006.
54. Wang, P. *Non-Axiomatic Logic: A Model of Intelligent Reasoning*; World Scientific: Singapore, 2013.
55. Capgemini Digital Transformation Institute. Understanding Digital Mastery Today. 2018. Available online: https://www.capgemini.com/wp-content/uploads/2018/07/Digital-Mastery-DTI-report_20180704_web.pdf (accessed on 19 December 2018).
56. Flach, P. *Machine Learning: The Art and Science of Algorithms That Make Sense of Data*; Cambridge University Press: New York, NY, USA, 2012.
57. Chomsky, N. *Language and Problems of Knowledge: The Managua Lectures*; MIT Press: Cambridge, MA, USA; London, UK, 1988.

information

MDPI

Article

Conceptions of Artificial Intelligence and Singularity

Pei Wang [1,*], Kai Liu [2] and Quinn Dougherty [3]

[1] Department of Computer and Information Sciences, Temple University, Philadelphia, PA 19122, USA
[2] School of Psychology, Central China Normal University, Wuhan 430079, China; ccnulk@mail.ccnu.edu.cn
[3] Philly AGI Team, Philadelphia, PA 19122, USA; quinn.dougherty.phila@gmail.com
[*] Correspondence: pei.wang@temple.edu

Received: 15 February 2018; Accepted: 3 April 2018; Published: 6 April 2018

Abstract: In the current discussions about "artificial intelligence" (AI) and "singularity", both labels are used with several very different senses, and the confusion among these senses is the root of many disagreements. Similarly, although "artificial general intelligence" (AGI) has become a widely used term in the related discussions, many people are not really familiar with this research, including its aim and status. We analyze these notions, and introduce the results of our own AGI research. Our main conclusions are that: (1) it is possible to build a computer system that follows the same laws of thought and shows similar properties as the human mind, but, since such an AGI will have neither a human body nor human experience, it will not behave exactly like a human, nor will it be "smarter than a human" on all tasks; and (2) since the development of an AGI requires a reasonably good understanding of the general mechanism of intelligence, the system's behaviors will still be understandable and predictable in principle. Therefore, the success of AGI will not necessarily lead to a singularity beyond which the future becomes completely incomprehensible and uncontrollable.

Keywords: artificial general intelligence; technical singularity; non-axiomatic reasoning system

1. Introduction

Driven by the remarkable achievements of deep learning, it becomes a hot topic again to debate whether computers can be smarter than humans. In the debate, there are two opposite tendencies that are both wrong in our opinion:

- Many claims about what AI (artificial intelligence) will be able to do are obtained using naive extrapolation of the past progress, without addressing the conceptual and technical difficulties in the field.
- Many claims about what AI will not be able to do are derived from traditional conceptions on how a computer system should be built and used, as well as an anthropocentric usage of notions such as "intelligence" and "cognition".

We consider the leading article of this Special Issue [1] as having done a good job in criticizing the first tendency by pointing out a list of features that any truly intelligent system should have, and arguing that mainstream AI techniques cannot deliver them, even after more research. However, many of the authors' conclusions exactly fall into the second tendency mentioned above, mainly because they are not familiar with existing AGI (artificial general intelligence) research. Since the opinions expressed in the leading article are representative, in this article, we will focus on the issues they raised, without addressing many related topics.

In the following, we start by distinguishing and clarifying the different interpretations and understandings of AI and singularity, and then explain how AGI is related to them. After that, we briefly summarize the AGI project our team has been working on, and explain how it can produce the features that the leading article claimed to be impossible for AI. In the conclusion, we agree with the authors of the leading article [1] that the recent achievements of deep learning are still far from showing that the related techniques can give us AGI or singularity; however, we believe AGI can be achieved via paths outside the vision of mainstream AI researchers, as well as that of its critics. This conception of AGI is fundamentally different from that of the current mainstream conception of AI. As for "singularity", we consider it an ill-conceived notion, as it is based on an improper conception of intelligence.

2. Notions Distinguished and Clarified

Let us first analyze what people mean when talking about "AI" and "Singularity". Both notions have no widely accepted definitions, although there are common usages.

2.1. Different Types of AI

In its broadest sense, AI is the attempt "to make a computer work like a human mind". Although it sounds plain, this description demands an AI to be similar (or even identical) to the human mind in certain aspects. On the other hand, because a computer is not a biological organism, nor does it live a human life, it cannot be expected to be similar to the human mind in all details. The latter is rarely mentioned but implicitly assumed, as it is self-evident. Consequently, by focusing on different aspects of the human mind, different paradigms of AI have been proposed and followed, with different objectives, desiderata, assumptions, road-maps, and applicabilities. They are each valid but distinct paradigms of scientific research [2].

In the current discussion, there are at least three senses of "AI" involved:

1. A computer system that behaves exactly like a human mind
2. A computer system that solves certain problems previously solvable only by the human mind
3. A computer system with the same cognitive functions as the human mind

In the following, they are referred to as AI-1, AI-2, and AI-3, respectively.

The best-known form of AI-1 is a computer system that can pass the Turing Test [3]. This notion is easy to understand and has been popularized by science-fiction novels and movies. To the general public, this is what "AI" means; however, it is rarely the research objective in the field, for several reasons.

At the very beginning of AI research, most researchers did attempt to build "thinking machines" with capabilities comparable (if not identical) to that of the human mind [3–5]. However, all direct attempts toward such goals failed [6–8]. Consequently, the mainstream AI researchers reinterpreted "AI" as AI-2, with a limited scope on a specific application or a single cognitive function. Almost all results summarized in the common AI textbooks [9,10] belong to this category, including deep learning [11] and other machine learning algorithms [12].

Although research on AI-2 has made impressive achievements, many people (both within the field and outside it) still have the feeling that this type of computer system is closer to traditional computing than to true intelligence, which should be general-purpose. This is why a new label, "AGI", was introduced more than a decade ago [13,14], even though this type of research projects has existed for many years. What distinguishes AGI from mainstream AI is that the former treats "intelligence" as one capability, while the latter treats it as a collection of loosely related capabilities. Therefore, AGI is basically the AI-3 listed above.

The commonly used phrase "Strong AI" roughly refers to AI-1 and AI-3 (AGI), in contrast to "Weak AI", referring to AI-2. Although this usage has intuitive attraction with respect to the ambition of the objectives, many AGI researchers usually do not use these phrases themselves, partly to avoid the philosophical presumptions behind the phrases [15]. Another reason is that the major difference between AI-2 and AI-3 is not in "strength in capability", but "breadth of applicability". For one concrete problem, a specially designed solution is often better than the solution provided by an AGI. We cannot expect an AI-2 technique which becomes "stronger" to eventually become AI-3, as the two are designed under fundamentally different considerations. For the same reason, it is unreasonable to expect to obtain an AI-3 system by simply bundling the existing AI-2 techniques together.

Furthermore, "Strong AI" fails to distinguish AI-1 and AI-3, where AI-1 focuses on the external behaviors of a system, while AI-3 focuses on its internal functions. It can be argued that "a computer system that behaves exactly like a human mind" (AI-1) may have to be built "with the same cognitive functions as the human mind" (AI-3); even so, the reverse implication is not necessarily true because the behaviors of a system, or its "output", not only depends on the system's processing mechanism and functions, but also on its "input", which can be roughly called the system's "experience". In the same way, two mathematical functions which are very similar may still produce very different output values if their input values are different enough [2].

In that case, why not give AGI human experience? In principle, it can be assumed that human sensory and perceptive processes can be simulated in computing devices to any desired accuracy. However, this approach has several obstacles. First, accuracy with regard to "human" sensory processes is not a trivial consideration. Take vision as an example: light sensors should have identical sensibility, resolution, response time, etc., as the human eye. That is much more to ask than for the computer to have "vision". Instead, it is to ask the computer to have "human vision", which is a special type of vision.

Even if we can simulate all human senses to arbitrary accuracy, they still can only produce the direct or physical experience of a normal human, but not the indirect or social experience obtained through communication, which requires the computer to be treated by others (humans and machines) as a human. This is not a technical problem, as many human beings will have no reason to do so.

For the sake of argument, let us assume the whole society indeed treats AGI systems exactly as if they were humans; in this case, AI-1 is possible. However, such an AI-1 is based on a highly anthropocentric interpretation of "intelligence", thus it should be called "Artificial Human Intelligence". To define general intelligence using human behavior would make other forms of intelligence (such as "animal intelligence", "collective intelligence", "extraterrestrial intelligence", etc.) impossible by definition, simply because they cannot have human-like inputs and outputs.

Such an anthropocentric interpretation of "intelligence" is rarely stated explicitly, although it is often assumed implicitly. One example is to take Turing Test as a working definition of AI, even though Turing himself only proposed it as a sufficient condition, but not a necessary condition, of intelligence or thinking. Turing [3] wrote: "May not machines carry out something which ought to be described as thinking but which is very different from what a man does? This objection is a very strong one, but at least we can say that if, nevertheless, a machine can be constructed to play the imitation game satisfactorily, we need not be troubled by this objection."

Among the current AGI researchers, we do not know anyone whose goal is to build an AI-1; instead, it is more proper to see their works as aiming at some version of AI-3. They believe "thinking machines" or "general intelligence" can be built which are comparable, or even identical to the human mind at a certain level of description, although not in all details of behaviors. These differences nevertheless do not disqualify these systems from being considered as truly intelligent, just like we consider fish and birds as having vision, despite knowing that what they see is very different from what we see.

What is currently called "AGI" is very similar to the initial attempts to do this under the name of "AI", and the new label was adopted around 2005 by a group of researchers who wanted to distinguish their objective from what was called "AI" at the time (i.e., AI-2). Since then, the AGI community has been mostly identified by its annual conference (started in 2008) and its journal (launched in 2009). Although there has been a substantial literature, as well as open-source projects, AGI research is still far from its goal; there is no widely accepted theory or model yet, not to mention practical application. As AGI projects typically take approaches unfavored by the mainstream AI community, the AGI community is still on the fringe of the field of AI, with their results largely unknown to the outside world, even though the term "AGI" has become more widely used in recent years.

On this aspect, the lead article [1] provides a typical example. Its main conclusion is that "strong AI" and "AGI" (the two are treated as synonymy) are impossible, and the phrase of "AGI" is used many times in the article, but its 68 references do not include even a single paper from the AGI conferences or journal, nor does the article discuss any of the active AGI projects, where most of the "ignored characteristics" claimed by Braga and Logan [1] have been explored, and demonstrable (although often preliminary) results have been produced. Here, we are not saying that AGI research cannot be criticized by people outside the field, but that such criticism should be based on some basic knowledge about the current status of the field.

In our opinion, one major issue of the lead article [1] is the failure to properly distinguish interpretations and understandings of "AI". We actually agree with its authors' criticism of mainstream AI and its associated hype, as well as the list of characteristics ignored in those systems. However, their criticism of AGI research is attacking a straw man, as it misunderstands AGI's objective (they assume it is that of AI-1) and current status (they assume it is that of AI-2).

2.2. Presumptions of Singularity

The "Singularity", also known as the "Technological Singularity", is another concept that has no accurate and widely accepted definition. It has not been taken to be a scientific or technical term, even though it has become well-known due to some writings for the general public (e.g., [16]).

In its typical usage, the belief that "AI will lead to singularity" can be analyzed into the conjunction of the following statements:

1. The intelligence of a system can be measured by a real number.
2. AI should be able to increase its intelligence via learning or recursive self-improvement.
3. After the intelligence of AI passes the human-level, its entire future will be perceived as a single point, since it will be beyond our comprehension.

However, some people also use "singularity" for the time when "human-level AI is achieved", or "computers have become more intelligent than human", without the other presumptions. In the following, we focus on the full version, although what we think about its variants should be quite clear after this analysis.

The first statement looks agreeable intuitively. After all, an "intelligent" or "smart" system should be able to solve many problems, and we often use various tests and examinations to evaluate that. In particular, human intelligence is commonly measured by an "intelligence quotient" (IQ). To accurately define a measurement of problem-solving capability for general-purpose systems will not be easy, but, for the sake of the current discussion, we assume such a measurement S can be established. Even so, we do not consider S a proper measurement of a system's "intelligence", as it misses the time component. In its common usage, the notion of intelligence is associated closer to "learned problem-solving capability" than to "innate problem-solving capability". For this reason, at a given time t, the intelligence of the system probably should not be measured by $S(t)$, but $S'(t)$, i.e., the increasing rate of S at the moment.

To visualize the difference between the two measurements, in Figure 1, there are four functions indicating how a system's total problem-solving score S is related to time t:

- **B-type:** The blue line corresponds to a system with constant problem-solving capability—what the system can do is completely determined at the beginning, i.e., $S'(t) = 0$. All traditional computation systems belong to this type, and some of them are referred to as "AI".
- **P-type:** The pink line corresponds to a system that increases its problem-solving capability until it infinitely approximates an upper bound. For such a system, $S'(t) > 0$, but converges to 0. Most machine learning algorithms belong to this type.
- **G-type:** The green line corresponds to a system where its problem-solving capability $S(t)$ increases without an upper bound and $S'(t)$ is a positive constant. Many AGI projects, including ours, belong to this type.
- **R-type:** The red line corresponds to a system where both $S(t)$ and $S'(t)$ increase exponentially. We do not think such a system is possible to be actually built, but list it here only as a conceptual possibility to be discussed.

Here, $S(t)$ is "problem-solving capability", while $S'(t)$ is "learning capability", and the two are not directly correlated in their values. As can be seen in Figure 1, depending on the constants and moment of measuring, each of the four types can be the most capable one in problem solving, but with respect to learning their order is basically the order of the previous descriptions: B < P < G < R.

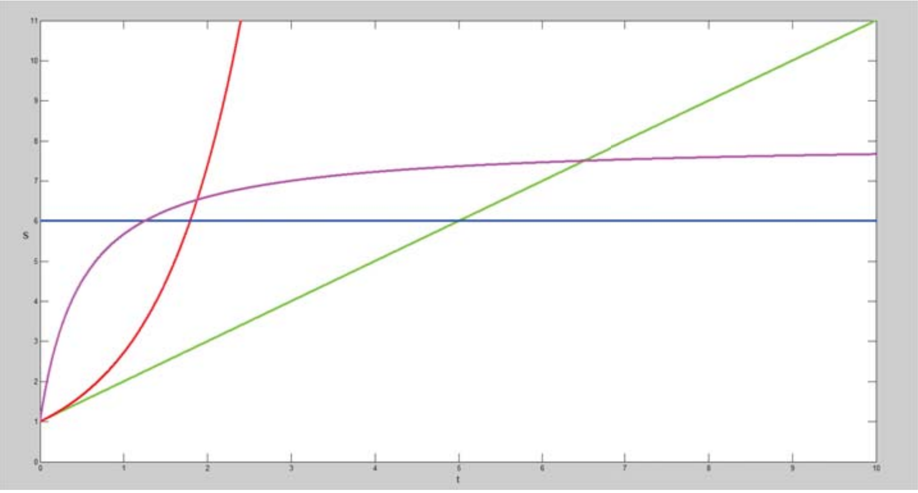


Figure 1. Four typical relations between time t and score S.

Our opinion is that "intelligence" should be measured by $S'(t)$, not $S(t)$. We believe this understanding of intelligence fits the deep sense of the word better, and will be more fruitful when used to guide the research of AI, although we know that it differs from current mainstream opinion.

The above conclusion does not necessarily conflict with the practice of human intelligence quotient (IQ) evaluation, which practically measures certain problem-solving capability. As IQ is the quotient obtained by dividing a person's "mental age" (according to the test score) by the person's chronological age. It can be interpreted as indicating the person's learning rate compared with that of other people, as higher $S(t)$ value implies higher $S'(t)$ value, given that the innate problem-solving capability $S(0)$ is not that different among human beings. However, this justification cannot be applied to AI systems, since they can have very different $S(0)$ values.

To place learning at the center of intelligence is not a new opinion at all, although usually learning is put on the same level as problem-solving. In our analysis, learning capability is at the meta-level, while the various problem-solving capabilities are at the object-level. This difference provides another perspective on the "AI vs. AGI" contrast mentioned previously. Mainstream AI research takes intelligence as the ability of solving specific problems, and for each problem its solution depends on problem-specific features. AGI, on the contrary, focuses on the meta-problems, which are independent of the specific domain. In this way, the two approaches actually do not overlap or compete, but complement each other.

For G-Type systems, it makes the discussion more clear by calling its meta-level knowledge and procedures "intelligence" (which is mostly built-in and independent of the system's experience), while calling its object-level knowledge and procedures "beliefs and skills" (which are mostly acquired from the system's experience). When we say such a system has reached "human-level", we mean its meta-level knowledge and procedures resemble those of the human mind, although its object-level beliefs and skills can overlap with that of a human being to an arbitrary extent [2].

The learning of meta-level knowledge in an AI system is possible, but there are several major issues that are rarely touched on in the relevant discussions:

- Although the distinction between object-level and meta-level exists in many AI systems, its exact standard depends on the design of the system, such that the functions carried out as meta-learning in one system may correspond to a type of object-level learning in another system.
- Some proposed "meta-learning" approaches basically suggest trying various designs and keeping the best, but this is usually unrealistic because of the time–space resources it demands and the risks it brings.
- A common misunderstanding is to equalize "recursive self-improving" with "modifying the system's own source-code". Certain programming languages, such as Lisp and Prolog, have provided the source-code modification and generation functions for many years; however, the use of these functions does not cause a fundamental difference, as the same effects can be achieved in another programming language by changing certain data, as the "data vs. code" distinction is also based on implementation considerations, not a theoretical boundary.
- The learning of meta-level knowledge and skills cannot be properly handled by the existing machine learning techniques, which are designed for object-level tasks [12]. The same difference happens in the human mind: as an individual, our mental mechanisms are mostly innate, and can be adjusted slowly in limited ways; it is only when discussed at the scale of the species that they are "learned" via evolution, but that happens at a much slower speed and a much higher price (a bad change is usually fatal for an individual). Given this fundamental difference, it is better to call the meta-level learning processes in a species "evolution", reserving the word "intelligence" for the object-level learning processes of an individual.

Thus far, we have not seen any convincing evidence for the possibility of a R-Type system. Although the existence of "exponential growth" is often claimed by the supporters of singularity, its evidence is never about the capability of a single system achieved by self-improvement. Although "intelligence" is surely a matter of degree, there is no evidence that the "level of intelligence" is an endless ladder with many steps above the "human-level". The existence of "super-intelligence" [17] is often argued using analogy from the existence of intelligence below the human-level (while mixing the object-level improvement with the meta-level improvement). Here, the situation is different from the above $S(t)$ values, which obviously can be increased from any point by adding knowledge and skills, as well as computational resources. At the meta-level, "above human" should mean a completely different thinking mechanism which better serves the purpose of adaptation. Of course, we cannot deny such a possibility, but have not seen any solid evidence.

Our hypothesis here is that the measurement of intelligence is just like a membership function of a fuzzy concept, with an upper-bound not much higher than the human-level. Furthermore, we believe it is possible for AGI to reach the same level; that is to say, there is a general notion of "intelligence" which can be understood and reproduced in computers. After that, the systems can still be improved, either by humans or by themselves, although there will not be a "super-intelligence" based on a fundamentally different mechanism.

Can a computer become smarter than a human? Sure, as this has already happened in many domains, if "smarter" means having a higher problem-solving capability. Computers have done better than human beings on many problems, but this result alone is not enough to earn them the title "intelligence", otherwise arithmetic calculators and sorting algorithms should also be considered as intelligent in their domains. To trivialize the notion of "intelligence" in this way will only lead to the need for a new notion to indicate the G-type systems. In fact, this is exactly why the phrase "AGI" was introduced. Similarly, the literal meaning of "machine learning" covers the G-type systems well, and the field of machine learning was also highly diverse at the beginning; however, now the phrase "machine learning" is usually interpreted as "function approximation using statistics", which only focuses on the P-type systems, so we have to use a different name to avoid misunderstanding [12,18].

Since, in a G-Type or R-Type system, $S(t)$ can grow to an arbitrary level, they can be "smarter than humans"; however, this does not mean that AGI will do better on every problem, usually for reasons such as sensor, actuator, or experience. On this matter, there is a fundamental difference between the G-Type systems and the R-Type systems: for the former, since the meta-level knowledge remains specified by its designer, we still understand how the system works in principle, even after its $S(t)$ value is far higher than what can be reached by a human being. On the contrary, if there were really such a thing as an R-Type system, it would reach a point beyond which we cannot even understand how it works.

Since we do not believe an R-Type system can exist, we do not think "singularity" (in its original sense) can happen. However, we do believe AGI systems can be built with meta-level capability comparable to that of a human mind (i.e., neither higher nor lower, although not necessarily identical), and object-level capability higher than that of a human mind (i.e., in total score, although not on every task). These two beliefs do not contradict each other. Therefore, although we agree with Braga and Logan [1] on the impossibility of a "singularity", our reasons are completely different.

3. What an AGI Can Do

To support our conclusions in the previous section, here we briefly introduce our own AGI project, NARS (Non-Axiomatic Reasoning System). First, we roughly describe how the system works, and then explain how the features listed by Braga and Logan [1] as essential for intelligence are produced in NARS.

The design of NARS has been described in two research monographs [19,20] and more than 60 papers, most of which can be downloaded at https://cis.temple.edu/~pwang/papers.html. In 2008, the project became open source, and since then has had more than 20 releases. The current version can be downloaded with documents and working examples at http://opennars.github.io/opennars/.

Given the complexity of NARS, as well as the nature and length of this article, here we merely summarize the major ideas in NARS' design in a non-technical language. Everything mentioned in the following about NARS has been implemented in computer, and described in detail in the aforementioned publications.

3.1. NARS Overview

Our research is guided by the belief that knowledge about human intelligence (HI) can be generalized into a theory on intelligence in general (GI), which can be implemented in a computer to become computer intelligence (CI, also known as AI), in that it keeps the cognitive features of HI, but without its biological features [21]. In this way, CI is neither a perfect duplicate nor a cheap substitute of HI, but is "parallel" to it as different forms of intelligence.

On how CI should be similar to HI, mainstream AI research focuses on what problems the system can solve, while our focus is on what problems the system can learn to solve. We do not see intelligence as a special type of computation, but as its antithesis, in the sense that "computation" is about repetitive procedures in problem solving, where the system has sufficient knowledge (an applicable algorithm for the problem) and resources (computational time and space required by the algorithm); "intelligence" is about adaptive procedures in problem solving, where the system has insufficient knowledge (no applicable algorithm) and resources (shortage of computational time and/or space) [22].

Based on such a belief, NARS is not established on the theoretical foundations of mainstream AI research (which mainly consist of mathematical logic, probability theory and the theory of computability and computational complexity), but on a theory of intelligence in which the Assumption of Insufficient Knowledge and Resources (hereafter AIKR) is taken as a fundamental constraint to be respected rigorously. Under AIKR, an adaptive system cannot merely execute the programs provided by its human designers, but must use its past experience to predict the future (although the past and the future are surely different), and use its available resources (supply) to best satisfy the pending demands (although the supply is always less than the demand).

To realize the above ideas in a computer system, NARS is designed as a reasoning system to simulate the human mind at the conceptual level, rather than at the neural level, meaning that the system's internal processing can be described as inference about conceptual relations.

Roughly speaking, the system's memory is a conceptual network, with interconnected concepts each identified by an internal name called a "term". In its simplest form, a term is just a unique identifier, or label, of a concept. To make the discussion natural, English nouns such as "bird" and "robin" are often used to name the terms in examples. A conceptual relation in NARS is taken to be a "statement", and its most basic type is called "inheritance", indicating a specialization-generalization relation between the terms and concepts involved. For example, the statement "$robin \to bird$" roughly expresses "Robin is a type of bird".

NARS is a reasoning system that uses a formal language, Narsese, for knowledge representation, and has a set of formal inference rules. Even so, it is fundamentally different from the traditional "symbolic" AI systems in several key aspects.

One such aspect is semantics, i.e., the definition of meaning and truth. Although the Narsese term *bird* intuitively corresponds to the English word "bird", the meaning of the former is not "all the birds in the world", but rather what the system already knows about the term at the moment according to its experience, which is a stream of input conceptual relations. Similarly, the truth-value of "$robin \to bird$" is not decided according to whether robins are birds in the real world, but rather the extent to which the term *robin* and the term *bird* have the same relations with other terms, according to evidence collected from the system's experience. For a given statement, available evidence can be either positive (affirmative) or negative (dissenting), and the system is always open to new evidence in the future.

A statement's truth-value is a pair of real numbers, both in [0, 1], representing the evidential support a statement obtains. The first number is "frequency", defined as the proportion of positive evidence to all available evidence. The second number is "confidence", defined as the proportion of currently available evidence to all projected available evidence at a future moment, after a new, constant amount of evidence is collected. Defined in this way, *frequency* is similar to probability, although it is only based on past observation and can change over time. *Confidence* starts at 0 (completely unknown) and gradually increases as new evidence is collected, but will never reach its upper-bound, 1 (completely known). NARS never treats an empirical statement as an axiom or absolute truth with a truth-value immune from future modification, which is why it is "non-axiomatic".

This "experience-grounded semantics" [23] of NARS bases the terms and statements of NARS directly on its experience, i.e., the system's record of its interaction with the outside world, without a human interpreter deciding meaning and truth. The system's beliefs are summaries of its experience, not descriptions of the world as it is. What a concept means to the system is determined by the role it plays in the system's experience, as well as by the attention the system currently pays to the concept, because under AIKR, when a concept is used, the system never takes all of its known relations into account. As there is no need for an "interpretation" provided by an observer, NARS cannot be challenged by Searle's "Chinese Room" argument as "only having syntax, but no semantics" [15,23].

In each inference step, NARS typically takes two statements with a shared term as its premises, and derives some conclusions according to the evidence provided by the premises. The basic inference rules are syllogistic, whose sample use-cases are given in Table 1.

Table 1. Sample steps of basic syllogistic inference.

Type	Deduction	Induction	Abduction
Premise 1	$robin \to bird$	$robin \to bird$	$robin \to [flyable]$
Premise 2	$bird \to [flyable]$	$robin \to [flyable]$	$bird \to [flyable]$
Conclusion	$robin \to [flyable]$	$bird \to [flyable]$	$robin \to bird$

The table includes three cases involving the same group of statements, where "$robin \to bird$" expresses "Robin is a type of bird", "$bird \to [flyable]$" expresses "Bird can fly", and "$robin \to [flyable]$" expresses "Robin can fly". For complete specification of Narsese grammar, see [20].

Deduction in NARS is based on the transitivity of the *inheritance* relation, that is, "if A is a type of B, and B is a type of C, then A is a type of C." This rule looks straightforward, except that since the two premises are true to differing degrees, so is the conclusion. Therefore, a truth-value function is part of the rule, which uses the truth-values of the premises to calculate the truth-value of the conclusion [20].

The other cases are induction and abduction. In NARS, they are specified as "reversed deduction" as in [24], obtained by switching the conclusion in deduction with one of the two premises, respectively. Without the associated truth-values, induction and abduction look unjustifiable, but according to experience-grounded semantics, in both cases the conclusion may get evidential support from the premise. Since each step only provides one piece of evidence, inductive and abductive conclusions normally have lower confidence than deductive conclusions.

NARS has a revision rule which merges evidence from distinct sources for the same statement, so the confidence of its conclusion is higher than that of the premises. Revision can also combine conclusions from different types of inference, as well as resolve contradictions by balancing positive and negative evidence.

To recognize complicated patterns in experience, Narsese has compound terms that each are constructed from some component terms, and NARS has inference rules to process these compounds. Certain terms are associated with the operations of sensors and actuators, therefore the system can represent procedural knowledge on how to do things, rather than just to talk about them. The grammar rules, semantic theory, and the inference rules altogether form the Non-Axiomatic Logic (NAL), the logic part of NARS [19,20].

From an user's point of view, NARS can accept three types of task:

- **Judgment:** A judgment is a statement with a given truth-value, as a piece of new knowledge to be absorbed into the system's beliefs. The system may revise the truth-value of a judgment according to its previous belief on the matter, add it into the conceptual network, and carry out spontaneous inference from it and the relevant beliefs to reveal its implications, recursively.
- **Goal:** A goal is a statement to be realized by the system. To indicate the extent of preference when competing with other goals, an initial "desire-value" can be given. When the desire-value of a goal becomes high enough, it will either directly trigger the execution of the associated operation, or generate derived goals according to the relevant beliefs.
- **Question:** A question can ask the truth-value or desire-value of a statement, which may contain variable terms to be instantiated. A question can be directly answered by a matching belief or desire, or generate derived questions according to the relevant beliefs.

These tasks and the system's beliefs (judgments that are already integrated into the system's memory) are organized into concepts according to the terms appearing in them. For example, tasks and beliefs on statement "$robin \rightarrow bird$" are referred from concept *robin* and concept *bird*. Each task only directly interacts with (i.e., being used as premises with) beliefs within the same concept, so every inference step happens within a concept.

As the system usually does not have the processing time and storage space to carry out the inference for every task to its completion (by exhaustively interacting with all beliefs in the concept), each data item (task, belief, and concept) has a priority value associated to indicate its share in resource competition. These priorities can take user specified initial values, and then be adjusted by the system according to the feedback (such as the usefulness of a belief, etc.).

NARS runs by repeating the following working cycle:

1. Select a concept in the system's memory probabilistically, biased by the priority distributions among concepts. Every concept has a chance to be selected, although concepts with high priority have higher chances.
2. Select a task and a belief from the concept, also probabilistically as above. Since the two share the same term identifying the concept, they must be relevant in content.

3. Carry out a step of inference using the selected task and belief as premises. Based on the combination of their types, the corresponding inference rules will be triggered, which may provide immediate solutions to the task, or (more likely) derived tasks whose solutions will contribute to the solution of the original task.
4. Adjust the priority values of the processed items (belief, task, and concept) according to the available feedback from this inference step.
5. Process the new tasks by selectively adding them into the memory and/or reporting them to the user.

3.2. Properties of NARS

Although the above description of NARS is brief and informal, it still provides enough information for some special properties of the system to be explained. A large part of [1] is to list certain "essential elements of or conditions for human intelligence" and claim they cannot be produced in AI systems. In this subsection, we describe how the current implementation of NARS generates these features (marked using bold font), at least in their preliminary forms. As each of them typically has no widely accepted definition, our understanding and interpretation of it will be inevitably different from that of other people, although there should be enough resemblance for this discussion to be meaningful.

The claim "Computers, like abacuses and slide rules, only carry out operations their human operators/programmers ask them to do, and as such, they are extensions of the minds of their operators/programmers." [1] is a variant of the so-called "Lady Lovelace's Objection" analyzed and rejected by Turing [3]. To many traditional systems, this claim is valid, but it is no longer applicable to adaptive systems like NARS. In such a system, what will be done for a problem not only depends on the initial design of the system, but also on the system's **experience**, which is the history of the system's interaction with the environment. In this simple and even trivial sense, every system has an experience, but whether it is worth mentioning is a different matter.

If a problem is given to a traditional system, and after a while a solution is returned, then if the same problem is repeated, the solving process and the solution should be repeated exactly, as this is how "computation" is defined in theoretical computer science [25]. In NARS, since the processing of a task will more or less change the system's memory irreversibly, and the system is not reset to a unique initial state after solving each problem, a repeated task will (in principle) be processed via a more or less different path—the system may simply report the previous answer without redoing the processing. Furthermore, the co-existing problem-solving processes may change the memory to make some concepts more accessible to suggest a different solution that the system had not previously considered. For familiar problems, the system's processing usually becomes stable, although whether a new problem instance belongs to a known problem type is always an issue to be considered from time to time by the system, rather than taken for granted.

Therefore, to accurately predict how NARS will process a task, to know its design is not enough. For the same reason, it is no longer correct to see every problem as being solved by the designer, because given the same design and initial content in memory, different experiences will actually lead to very different systems, in terms of their concepts, beliefs, skills, etc. Given this situation, it makes more sense to see the problems as solved by the system itself, even though this **self** is not coming out of nowhere magically or mythically, but rooted in the initial configuration and shaped by the system's experience.

NARS has a $self$ concept as a focal point of the system's self-awareness and self-control. Like all concepts in NARS, the content of $self$ mainly comes from accumulated and summarized experience about the system itself, although this concept has special innate (built-in) relations with the system's primary operations. It means at the very beginning the system's "self" is determined by "what I can do" and "what I can feel" (since in NARS perception is a special type of operation), but gradually it will learn "what is my relation with the outside objects and systems", so the concept becomes more and more complicated [26]. Just like NARS' knowledge about the environment, its knowledge about itself is always uncertain and incomplete, but we cannot say that it has no sense of itself.

NARS can be equipped with various sensors, and each type of sensor expends the system's experience to a new dimension by adding a new sensory channel into the system where a certain type of signals are recognized, transformed into Narsese terms, then organized and generalized via a perceptive process to enter the system's memory. The sensors can be on either the external environment or the internal environment of the system, where the latter provides self-awareness about what has been going on within the system. Since the internal experience is limited to significant internal events only, in NARS the conscious/unconscious distinction can be meaningfully drawn, according to whether an internal event is registered in the system's experience and becomes a belief expressed in Narsese.

The interactions between unconscious and conscious mental events were argued to be important by Freud [27], and this opinion is supported by recent neuroscientific study [28]. As only significant events within NARS enter the system's (conscious) experience, the same conclusion holds for NARS. A common misunderstanding about NARS-like systems is that all events in such a system must be conscious to the system, or that the distinction between conscious and unconscious events is fixed. Neither is correct in NARS, mainly because of AIKR, as an event can change its conscious status merely because of its priority level adjustments [26]. This interaction in NARS has a purely functional explanation that has little dependency on the detail of human neural activities.

As far as the system can tell consciously, its decisions are made according to its own **free will**, rather than by someone else or according to certain predetermined procedures, simply because the system often has to deal with problems for which no ready made solutions are there, so it has to explore the alternatives and weigh the pros and cons when a decision is made, all by itself. For an omniscient observer, all the decisions are predetermined by all the relevant factors collectively, but even from that viewpoint, it is still the decision by the system, not by its designer, who cannot predetermine the experience factor.

Given the critical role played by experience, it is more natural to accredit certain responsibility and achievement to the system, rather than to the designer. The system's beliefs are not merely copies of what it was taught by the user, but summaries of its experience. These beliefs include moral **judgments** (beliefs about what are good and what are bad, according to its desires and goals), **wisdom** (beliefs that guide the system to achieve its goals), **intuition** (beliefs whose source is too complicated or vague to recall), and so on. These beliefs are often from the view point of the system as they are produced by its unique experience. Even so, the beliefs of NARS will not be purely subjective, as the system's communication with other systems provide social experience for it, and consequently the relevant beliefs will have certain objective (or more accurately, "intersubjective") flavors in it, in the sense that it is not fully determined by the system's idiosyncratic experience, but strongly influenced by the community, society, or culture that the system belongs to.

Not only should the beliefs in NARS be taken as "of the system's own", but also the **desires** and **goals**. The design of NARS does not presume any specific goal, so all the original goals come from the outside, that is, the designer or the user. NARS has a goal derivation mechanism that generates derived goals from the existing (input or derived) goals and the relevant beliefs. Under AIKR, a derived goal G_2 is treated independently of its "parent" goal G_1, so in certain situation it may become more influential than G_1, and can even suppress it. Therefore, NARS is not completely controlled by its given goals, but also by the other items in its experience, such as the beliefs on how the goals can be achieved. This property is at the core of autonomy, originality, and creativity, although at the same time it raises a challenge on how to make the system behave according to human interests [29].

As an AGI, the behavior of NARS is rarely determined by a single goal, but often by a large number of competing and even conflicting goals and desires. When an operation is executed, it is usually driven by the "resultant" of the whole motivation complex, rather than by one motivation [29]. This motivation complex develops over time, and also contributes greatly to the system's self identity.

In different contexts, we may describe the difference aspects of this complex as **purpose**, **objective**, **telos**, and even **caring**.

Desires and goals with special content are often labeled using special words. For example, when its social experience become rich and complicated enough, NARS may form what may be called "**values**" and "**morality**", as they are about how a system should behave when dealing with other systems. When the content of a goal drives the system to explore an unknown territory without explicitly specified purpose, we may call it "**curiosity**". However, the fact that we have a word for a phenomenon does not mean that it is produced by an independent mechanism. Instead, the phenomena discussed above are all generated by the same process in NARS, although each time we selectively discuss some aspects of it, or set up its context differently.

A large part of argument for the impossibility of AGI in [1] is organized around the "figure–ground" metaphor, where a key ingredient of the "ground" is **emotion**, which is claimed to be impossible in computers. However, this repeated claim only reveals the lack of knowledge of the authors about the current AGI research, as many AGI projects have emotion as a key component [30–32]. In the following, we only introduce the emotional mechanism in NARS, which is explained in detail in [33].

In NARS, emotion starts as an appraisal of the current situation, according to the system's desires. On each statement, there is a truth-value indicating the current situation, and a desire-value indicating what the system desires the situation to be, according to the relevant goals. The proximity of these two values measures the system's "satisfaction" on this matter. At the whole system level, there is an overall satisfaction variable that accumulates the individual measurements on the recently processed tasks, which will produce a positive or negative appraisal of the overall situation. That is, the system will have positive emotion if the reality agrees to its desires, and negative emotion if the reality disagrees to its desires.

These satisfaction values can be "felt" by the system's inner sensors, as well as be involved in the system's self-control. For instance, events associated with strong (positive or negative) emotion will get more attention (and therefore more processing resources) than the emotionally neutral events. When the system is in positive emotion, it is more acceptive to new tasks (meaning it devotes to them more resources). A strong emotion for someone or something corresponds to the phenomenon of "**passion**".

At the moment, we are extending the emotional mechanism in several ways, including to further distinguish different emotions (such as "**pleasure**" and "**joy**" at the positive side, and "scare" and "anger" at the negative side), to use emotion in communication, to control the effect of emotion in decision making, and so on.

Among the features in the list of [1], the only ones that have not been directly addressed in the previous publications and implementations of NARS are **imagination**, **aesthetics**, and **humor**. We do have plan to realize them in NARS, but will not discuss it in this article.

In summary, we agree the features listed in [1] are all necessary for AGI, and we also agree that the mainstream AI techniques cannot generate them. However, we disagree with the conclusion that they cannot be generated in computer systems at all. On the contrary, most of them have been realized in NARS in their preliminary form, and NARS is not the only AGI project that has addressed these topics.

Of course, we are not claiming that all these phenomena have been fully understood and perfectly reproduced in NARS or other AGI systems. On the contrary, the study of them is still in an early stage, and there are many open problems. However, the results so far have at least shown their possibility, or, more accurately, their inevitability, to appear in AGI systems. As shown above, in NARS these features are not added in one by one for their own sake, but are produced altogether from the design of NARS, usually as implications of AIKR.

A predictable objection to our above conclusions is to consider the NARS versions of these features to be "fake", as they are not identical to the human versions here or there. Once again, it goes back to the understanding of "AI" and how close it should be to human intelligence. Take emotion as an example: even when fully developed, the emotions in NARS will not be identical to human emotions, nor will they be accompanied by the physiological processes that are intrinsic ingredients of human emotion. However, these differences cannot be used to judge emotions in AGI as fake, as long as "emotion" is taken as a generalization of "human emotion" by keeping the functional aspects but not the biological ones.

Here is what we see as our key difference with Braga and Logan [1]: while we fully acknowledge that the original and current usage of the features they listed are tied to the human mind/brain complex, we believe it is both valid and fruitful to generalize these concepts to cover non-human and even non-biological systems, as their core meaning is not biological, but functional. Such a belief is also shared by many other researchers in the field, although how to accurately define these features is still highly controversial.

4. Conclusions

In this article, we summarize our opinions on AI, AGI, and singularity, and use our own AGI system as evidence to support these opinions. The purpose of this article is not to introduce new technical ideas, as the aspects of NARS mentioned above have all been described in our previous publications. Since many people are not familiar with the results of AGI research (as shown in this Special Issue of *Information*), we consider it necessary to introduce them to clarify the relevant notions in the discussion on what can be achieved in AI systems.

We agree with the lead article [1] that the mainstream AI techniques will not lead to "Strong AI" or AGI that is comparable to human intelligence in general, or to a "Singularity" where AI becomes "smarter than human", partly because these techniques fail to reproduce a group of essential characteristics of intelligence.

However, we disagree with their conclusion that AGI is completely impossible because the human mind is fundamentally different from digital computers [1], partly because most of the characteristics they listed have already been partially realized in our system. In our opinion, there are the following major issues in their argument:

- Their understanding of "intelligence" is highly anthropocentric. As few AGI researcher aims at such an objective, the "AGI research" they criticize does not exist [2].
- Their understanding of computer is oversimplified and reductionist, and only corresponds to a special way of using computer. Even though this is indeed the most common way for a computer system to be built and used at the present time, it is not the only possible way [34].
- Their understanding of AGI mostly comes from outsiders' speculations, rather than from the actual research activity in the field. Although it is perfectly fine for an outsider to criticize AGI research, such a criticism is valid only when it is based on the reality of the field.

As with respect to the topics under discussion, our positions are:

- AGI should be conceived as a computer system that is similar to human intelligence in principles, mechanisms, and functions, but not necessarily in internal structure, external behaviors, or problem-solving capabilities. Consequently, as another form of intelligence, AGI will roughly be at the same level of competence as human intelligence, neither higher nor lower. As in concrete problem-solving capability, AGI is not always comparable to human intelligence, since they may deal with different problems in different environments.
- To achieve AGI, new theories, models, and techniques are needed. The current mainstream AI results will not naturally grow in this direction, because they are mainly developed according to the dogma that intelligence is problem-solving capability, which does not correspond to AGI, but a fundamentally different objective, with different theoretical and practical values.
- Even when AGI is achieved, it does not lead to a singularity beyond which intelligent computer systems become completely incomprehensible, unpredictable, and uncontrollable. On the contrary, the achieving of AGI means the essence of intelligence has been captured by humans, which will further guide the use of AGI to meet to human values and needs.

AGI research is still in an early stage, and opinions from all perspectives are valuable, although it is necessary to clarify the basic notions to set up a minimum common ground, so the voices will not talk past each other. For this reason, the current Special Issue of *Information* is a valuable effort.

Author Contributions: P.W. conceived the article and submitted an abstract; P.W. drafted the article after discussions with K.L. and Q.D.; K.L. and Q.D. revised the draft; Q.D. corrected the English.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Braga, A.; Logan, R.K. The Emperor of Strong AI Has No Clothes: Limits to Artificial Intelligence. *Information* **2017**, *8*, 156, doi:10.3390/info8040156.
2. Wang, P. What do you mean by "AI". In Proceedings of the First Conference on Artificial General Intelligence, Memphis, TN, USA, 1–3 March 2008; pp. 362–373.
3. Turing, A.M. Computing machinery and intelligence. *Mind* **1950**, *LIX*, 433–460.
4. McCarthy, J.; Minsky, M.; Rochester, N.; Shannon, C. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, 1955. Available online: http://www-formal.stanford.edu/jmc/history/dartmouth.html (accessed on 4 April 2018).
5. Feigenbaum, E.A.; Feldman, J. *Computers and Thought*; McGraw-Hill: New York, NY, USA, 1963.
6. Newell, A.; Simon, H.A. GPS, a program that simulates human thought. In *Computers and Thought*; Feigenbaum, E.A., Feldman, J., Eds.; McGraw-Hill: New York, NY, USA, 1963; pp. 279–293.
7. Fuchi, K. The significance of fifth-generation computer systems. In *The Age of Intelligent Machines*; Kurzweil, R., Ed.; MIT Press: Cambridge, MA, USA, 1990; pp. 336–345.
8. Roland, A.; Shiman, P. *Strategic Computing: DARPA and the Quest for Machine Intelligence, 1983–1993*; MIT Press: Cambridge, MA, USA, 2002.
9. Luger, G.F. *Artificial Intelligence: Structures and Strategies for Complex Problem Solving*, 6th ed.; Pearson: Boston, MA, USA, 2008.
10. Russell, S.; Norvig, P. *Artificial Intelligence: A Modern Approach*, 3rd ed.; Prentice Hall: Upper Saddle River, NJ, USA, 2010.
11. LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436–444, doi:10.1038/nature14539.
12. Flach, P. *Machine Learning: The Art and Science of Algorithms That Make Sense of Data*; Cambridge University Press: New York, NY, USA, 2012.
13. Pennachin, C.; Goertzel, B. Contemporary approaches to artificial general intelligence. In *Artificial General Intelligence*; Goertzel, B., Pennachin, C., Eds.; Springer: New York, NY, USA, 2007; pp. 1–30.
14. Wang, P.; Goertzel, B. Introduction: Aspects of artificial general intelligence. In *Advance of Artificial General Intelligence*; Goertzel, B., Wang, P., Eds.; IOS Press: Amsterdam, The Netherlands, 2007; pp. 1–16.
15. Searle, J. Minds, brains, and programs. *Behav. Brain Sci.* **1980**, *3*, 417–424.
16. Kurzweil, R. *The Singularity Is Near: When Humans Transcend Biology*; Penguin Books: New York, NY, USA, 2006.
17. Bostrom, N. *Superintelligence: Paths, Dangers, Strategies*; Oxford University Press: Oxford, UK, 2014.
18. Wang, P.; Li, X. Different Conceptions of Learning: Function Approximation vs. Self-Organization. In Proceedings of the Ninth Conference on Artificial General Intelligence, New York, NY, USA, 16–19 July 2016; pp. 140–149.
19. Wang, P. *Rigid Flexibility: The Logic of Intelligence*; Springer: Dordrecht, The Netherlands, 2006.
20. Wang, P. *Non-Axiomatic Logic: A Model of Intelligent Reasoning*; World Scientific: Singapore, 2013.
21. Wang, P. Theories of Artificial Intelligence—Meta-theoretical considerations. In *Theoretical Foundations of Artificial General Intelligence*; Wang, P., Goertzel, B., Eds.; Atlantis Press: Paris, France, 2012; pp. 305–323.
22. Wang, P. The Assumptions on Knowledge and Resources in Models of Rationality. *Int. J. Mach. Conscious.* **2011**, *3*, 193–218.
23. Wang, P. Experience-grounded semantics: A theory for intelligent systems. *Cognit. Syst. Res.* **2005**, *6*, 282–302.
24. Peirce, C.S. *Collected Papers of Charles Sanders Peirce*; Harvard University Press: Cambridge, MA, USA, 1931; Volume 2.
25. Hopcroft, J.E.; Motwani, R.; Ullman, J.D. *Introduction to Automata Theory, Languages, and Computation*, 3rd ed.; Addison-Wesley: Boston, MA, USA, 2007.
26. Wang, P.; Li, X.; Hammer, P. Self in NARS, an AGI System. *Front. Robot. AI* **2018**, *5*, 20, doi:10.3389/frobt.2018.00020.

27. Freud, S. *The Interpretation of Dreams*; Translated by James Strachey from the 1900 Edition; Avon Books: New York, NY, USA, 1965.
28. Dresp-Langley, B. Why the Brain Knows More than We Do: Non-Conscious Representations and Their Role in the Construction of Conscious Experience. *Brain Sci.* **2012**, *2*, 1–21.
29. Wang, P. Motivation Management in AGI Systems. In Proceedings of the Fifth Conference on Artificial General Intelligence, Oxford, UK, 8–11 December 2012; pp. 352–361.
30. Bach, J. Modeling Motivation and the Emergence of Affect in a Cognitive Agent. In *Theoretical Foundations of Artificial General Intelligence*; Wang, P., Goertzel, B., Eds.; Atlantis Press: Paris, France, 2012; pp. 241–262.
31. Franklin, S.; Madl, T.; D'Mello, S.; Snaider, J. LIDA: A Systems-level Architecture for Cognition, Emotion, and Learning. *IEEE Trans. Auton. Ment. Dev.* **2014**, *6*, 19–41.
32. Rosenbloom, P.S.; Gratch, J.; Ustun, V. Towards Emotion in Sigma: From Appraisal to Attention. In Proceedings of the Eighth Conference on Artificial General Intelligence, Berlin, Germany, 22–25 July 2015; pp. 142–151.
33. Wang, P.; Talanov, M.; Hammer, P. The Emotional Mechanisms in NARS. In Proceedings of the Ninth Conference on Artificial General Intelligence, New York, NY, USA, 16–19 July 2016; pp. 150–159.
34. Wang, P. Three fundamental misconceptions of artificial intelligence. *J. Exp. Theor. Artif. Intell.* **2007**, *19*, 249–268.

information

Commentary

The Singularity May Be Near

Roman V. Yampolskiy

Department of Computer Engineering and Computer Science, Speed School of Engineering, University of Louisville, Louisville, KY 40292, USA; roman.yampolskiy@louisville.edu

Received: 5 July 2018; Accepted: 25 July 2018; Published: 27 July 2018

Abstract: Toby Walsh in "*The Singularity May Never Be Near*" gives six arguments to support his point of view that technological singularity may happen, but that it is unlikely. In this paper, we provide analysis of each one of his arguments and arrive at similar conclusions, but with more weight given to the "likely to happen" prediction.

Keywords: autogenous intelligence; bootstrap fallacy; recursive self-improvement; self-modifying software; singularity

1. Introduction

In February of 2016, Toby Walsh presented his paper "*The Singularity May Never Be Near*" at AAAI16 [1], which was archived on 20 February, 2016. In it, Walsh analyzes the concept of technological singularity. He does not argue that Artificial Intelligence (AI) will fail to achieve super-human intelligence; rather he suggests that it may not lead to the runaway exponential growth. Walsh notes that there is a lot of optimism and pessimism in the field of Artificial Intelligence (AI). Optimists are investing billions of dollars in AI. On the other hand, pessimists expect AI to end many things: jobs, wars, and even humanity. Both optimists and pessimists often turn to the idea of technological singularity: the time when AI will be able to take over AI research, and a new, much more intelligent species begins to populate the world. If optimists are right, it will be a moment that will fundamentally change our economy and society. If pessimists are right, it will also be a moment that will significantly change our economy and society. It is, therefore, worthwhile to invest some time to decide whether one of them may be right [1]. Walsh defends his view via six different arguments.

Almost exactly a year before, on 23 February 2015, Roman Yampolskiy archived his paper "*From Seed AI to Technological Singularity via Recursively Self-Improving Software*" [2] which was subsequently published as two peer-reviewed papers at AGI15 [3,4]. In it, Yampolskiy makes arguments similar to those made by Walsh, but also considers evidence in favor of intelligence explosion. Yampolskiy's conclusion is that singularity may not happen but leans more toward it happening. In the next section, we present arguments from the original paper by Yampolskiy mapped to each of the six arguments given by Walsh in his work.

2. Contrasting Yampolskiy's and Walsh's Arguments

To make it easier to contrast arguments derived from *On the Limits of Recursively Self-Improving Artificially Intelligent Systems* [2], we use Walsh's naming of arguments, even if our analysis does not rely on the same example (e.g., No dog).

2.1. Fast-Thinking Dog

Walsh argues: " … speed alone does not bring increased intelligence". Yampolskiy says: "In practice, the performance of almost any system can be trivially improved by the allocation of additional computational resources such as more memory, higher sensor resolution, faster processor, or greater network bandwidth for access to information. This linear scaling does not fit the definition of recursive improvement, as the system does not become better at improving itself. To fit the definition, the system would have to engineer a faster type of memory, not just purchase more memory units of the type it already has access to. In general, hardware improvements are likely to speed up the system, while software improvements (novel algorithms) are necessary for achievement of meta-improvements." It is clear from the original paper that performance in this context is the same as intelligence, and as most of our intelligence testing tools (IQ tests) are time-based, increased speed would in fact lead to higher Intelligence Quotient, at least in terms of how we currently access intelligence.

2.2. Anthropocentric

Walsh argues that "human intelligence is itself nothing special", meaning it is not a point that "once passed, allows for rapid increases in intelligence". Yampolskiy says: "We still do not know the minimum intelligence necessary for commencing the RSI (Recursive Self-Improvement) process, but we can argue that it would be on par with human intelligence, which we associate with universal or general intelligence [5], though in principal, a sub-human level system capable of self-improvement cannot be excluded [6]. One may argue that even human-level capability is not enough, because we already have programmers (people or their intellectual equivalence formalized as functions [7], or Human Oracles [8,9]) who have access to their own source code (DNA), but who fail to understand how DNA (nature) works to create their intelligence. This does not even include the additional complexity in trying to improve on existing DNA code or complicating factors presented by the impact of learning environment (nurture) on the development of human intelligence. Worse yet, it is not obvious how much above human ability an AI needs to be to begin overcoming the 'complexity barrier' associated with self-understanding."

2.3. Meta-Intelligence

Walsh argues: " . . . strongest arguments against the idea of a technological singularity in my view is that it confuses intelligence to do a task with the capability to improve your intelligence to do a task" and cites a quote from Chalmers [6] as an example: "If we produce an AI by machine learning, it is likely that soon after, we will be able to improve the learning algorithm and extend the learning process, leading to AI+". Yampolskiy says: "Chalmers [6] uses logic and mathematical induction to show that if an AI_0 system is capable of producing an only slightly more capable AI_1 system, a generalization of that process leads to a super-intelligent performance in AI_n after n generations. He articulates that his proof assumes that the *proportionality thesis,* which states that increases in intelligence lead to proportionate increases in the capacity to design future generations of AIs, is true."

2.4. Diminishing Returns

Walsh argues: "There is often lots of low hanging fruit at the start, but we then run into great difficulties to improve after this. . . . An AI system may be able to improve itself an infinite number of times, but the extent to which its intelligence changes overall could be bounded." Yampolskiy says, " . . . the law of diminishing returns quickly sets in, and after an initial significant improvement phase, characterized by discovery of 'low-hanging fruit', future improvements are likely to be less frequent and less significant, producing a bell curve of valuable changes."

2.5. Limits of Intelligence

Walsh argues: "There are many fundamental limits within the universe". Yampolskiy outlines such limits in great detail: "First of all, any implemented software system relies on hardware for memory, communication, and information processing needs, even if we assume that it will take a non-Von Neumann (quantum) architecture to run such software. This creates strict theoretical limits to computation, which despite hardware advances predicted by Moore's law will not be overcome by any future hardware paradigm. Bremermann [10], Bekenstein [11], Lloyd [12], Anders [13], Aaronson [14], Shannon [15], Krauss [16], and many others have investigated the ultimate limits to computation in terms of speed, communication, and energy consumption, with respect to such factors as speed of light, quantum noise, and gravitational constant." "In addition to limitations endemic to hardware, software-related limitations may present even bigger obstacles for RSI systems. Intelligence is not measured as a standalone value, but with respect to the problems it allows to solve. For many problems such as playing checkers [17], it is possible to completely solve the problem (provide an optimal solution after considering all possible options) after which no additional performance improvement would be possible [18]."

2.6. Computational Complexity

Walsh argues: " … no amount of growth in performance will make undecidable problems decidable" and Yampolskiy says, "Other problems are known to be unsolvable regardless of level of intelligence applied to them [19]. Assuming separation of complexity classes (such as P vs. NP) holds [20], it becomes obvious that certain classes of problems will always remain only approximately solvable and any improvements in solutions will come from additional hardware resources, not higher intelligence."

3. Response to Walsh's Arguments

In this section, we provide novel analysis of all six arguments presented by Walsh and, via mapping provided in the previous section, revisit and critically analyze arguments made by Yampolskiy.

3.1. Fast-Thinking Dog

This argument intuitively makes sense, since nobody has ever managed to train a dog to play chess. However, intuition is no match for a scientific experiment. Animals have successfully been trained to understand and even use human (sign) language and do some basic math. People with mental and learning disabilities, who have been long considered a "lost cause", have been successfully trained to perform very complex behaviors via alternative teaching methods and longer training spans. It is entirely possible that if one had thousands of years to train a dog, it would learn to play a decent game of chess; after all, it has a neural network very similar to the one used by humans and deep-learning AI. It may be argued that there is considerable evidence that language and other capabilities are functions of specific brain structures that are largely absent from a dog. Thus, thousands of years of training would not suffice, and one would need millions of years of evolution to get a human-level intelligent dog. However, some recent research has documented that people missing most of their brain can have near normal cognitive capacity [21], and even significant damage to parts of the brain can be overcome due to neuroplasticity [22], suggesting that brain structures are much more general. To transfer an analogy to another domain, an Intel286 processor is not fast enough to perform live speech recognition, but if you speed it up, it is. Until an actual experiment can be performed on an accurately simulated digital dog, this argument will remain speculative.

3.2. Anthropocentric

The reason some experts believe ([23], p. 339; [24], Chapter 3) that the human level of intelligence is special is not owing to an anthropocentric bias, but rather to the Church-Turing Thesis (CTT). The CTT states that a function over natural numbers is computable by a prototypical human being if and only if it is computable by some Turing Machine (TM), assuming that such a theoretical human has infinite computational resources similar to an infinite tape available to a TM. This creates an equivalence between human level intelligence and a Universal Turing Machine, which is a very special machine in terms of its capabilities. However, it is important to note that the debate regarding provability of the CTT remains open [25,26].

3.3. Meta-Intelligence

If the system is superior to human performance in all domains, as required by the definition of superintelligence, it would also be superior in the domain of engineering, computer science, and AI research. Potentially, it would be capable of improving intelligence of its successor up to any theoretical and physical limits, which might represent an upper bound on optimization power. In other words, if it were possible to improve intelligence, a super-intelligent system would do so; but as such possibility remains a speculation, this is probably the strongest of all presented objections to intelligence explosion.

3.4. Diminishing Returns

It is a mathematical fact that many functions, while providing diminishing returns, continue diverging. For example, the harmonic series $(1 + 1/2 + 1/3 + 1/4 + 1/5 + \ldots = \infty)$ is a highly counterintuitive result, yet is a proven mathematical fact. Additionally, as the system itself would be continuously improving, it is possible that the discoveries it would make with respect to future improvements would also improve in terms of their impact on the overall intelligence of the system. Thus, while it is possible that diminishing returns could be encountered, it is just as possible that returns would not be diminished.

3.5. Limits of Intelligence

While physical and theoretical limits to intelligence definitely exist, they may be far beyond our capacity to attain in practice, and so would have no impact on our perception of machine intelligence appearing to be undergoing intelligence explosion. It is also possible that physical constants are not permanently set, but dynamically changing, which has been demonstrated for some such physical "constants". It is also possible that the speed of improvement in intelligence would be below the speed with which some such constants would change. To bring an example from another domain, our universe can be said to be expanding faster than the speed of light, with respect to distance between some selected regions, so even with travel at a maximum theoretical speed (of light), we would never hit a threshold. Therefore, this is another open question, and a limit may or may not be encountered in the process of self-improvement.

3.6. Computational Complexity

While it is certainly true that undecidable problems remain undecidable, it is not a limitation on intelligence explosion, as it is not a requirement to qualify as super-intelligent. Moreover, plenty of solvable problems exists at all levels of difficulty. Walsh correctly points out that most limitations associated with computational complexity are only problems with our current models of computation and are avoided by switching to different paradigms of computation, such as quantum computing and some, perhaps not yet discovered, implementations of hypercomputation.

4. Conclusions

Careful side-by-side analysis of papers by Walsh and Yampolskiy shows an almost identical set of arguments against the possibility of technological singularity. This level of successful replication in analysis is an encouraging fact in science and gives additional weight to shared conclusions. Nevertheless, in this paper we provide a novel analysis of Walsh's/Yampolskiy's arguments which shows that they may not be as strong as they might initially appear. Future productive directions of analysis may concentrate on a number of inherent advantages, which may permit AI to recursively self-improve [27] and possibly succeed in this challenging domain: the ability to work uninterruptedly (no breaks, sleep, vocation, etc.), omniscience (complete and cross-disciplinary knowledge), greater speed and precision (brain vs. processor, human memory vs. computer memory), intersystem communication speed (chemical vs. electrical), duplicability (intelligent software can be copied), editability (source code unlike DNA can be quickly modified), near-optimal rationality (if not relying on heuristics) [28], advanced communication (ability to share cognitive representations complex concepts), new cognitive modalities (sensors for source code), ability to analyze low level hardware (e.g., individual registers), and addition of hardware (ability to add new memory, processors, etc.) [29]. The debate regarding possibility of technological singularity will continue. Interested readers are advised to read the full paper by Yampolskiy [2], as well as a number of excellent relevant chapters in Singularity Hypothesis [30] which address many arguments not considered in this paper.

Funding: This research received no external funding.

Acknowledgments: The author wishes to thank Toby Walsh for encouraging and supporting work on this paper, as well as reviewers who provided feedback on an early draft and by doing so, made the arguments presented in the paper much stronger.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Walsh, T. The Singularity May Never be Near. In Proceedings of the 2nd International Workshop on AI, Ethics and Society (AIEthicsSociety2016) & 30th AAAI Conference on Artificial Intelligence (AAAI-2016), Phoenix, AZ, USA, 12–13 February 2016.
2. Yampolskiy, R.V. From Seed AI to Technological Singularity via Recursively Self-Improving Software. *arXiv* **2015**, arXiv:1502.06512.
3. Yampolskiy, R.V. Analysis of Types of Self-Improving Software. In Proceedings of the Artificial General Intelligence: 8th International Conference (AGI 2015), Berlin, Germany, 22–25 July 2015; Volume 9205, p. 384.
4. Yampolskiy, R.V. On the Limits of Recursively Self-Improving AGI. In Proceedings of the Artificial General Intelligence: 8th International Conference (AGI 2015), Berlin, Germany, 22–25 July 2015; Volume 9205, p. 394.
5. Loosemore, R.; Goertzel, B. *Why an Intelligence Explosion Is Probable, Singularity Hypotheses*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 83–98.
6. Chalmers, D. The Singularity: A Philosophical Analysis. *J. Conscious. Stud.* **2010**, *17*, 7–65.
7. Shahaf, D.; Amir, E. Towards a theory of AI completeness. In Proceedings of the 8th International Symposium on Logical Formalizations of Commonsense Reasoning (Commonsense 2007), Stanford, CA, USA, 26–28 March 2007.
8. Yampolskiy, R. Turing Test as a Defining Feature of AI-Completeness. In *Artificial Intelligence, Evolutionary Computing and Metaheuristics*; Yang, X.-S., Ed.; Springer: Berlin/Heidelberg, Germany, 2013; pp. 3–17.
9. Yampolskiy, R.V. AI-Complete, AI-Hard, or AI-Easy–Classification of Problems in AI. In Proceedings of the 23rd Midwest Artificial Intelligence and Cognitive Science Conference, Cincinnati, OH, USA, 21–22 April 2012.
10. Bremermann, H.J. Quantum noise and information. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, USA, 27 December 1966–7 January 1967; pp. 15–22.
11. Bekenstein, J.D. Information in the holographic universe. *Sci. Am.* **2003**, *289*, 58–65. [CrossRef] [PubMed]
12. Lloyd, S. Ultimate Physical Limits to Computation. *Nature* **2000**, *406*, 1047–1054. [CrossRef] [PubMed]
13. Sandberg, A. The physics of information processing superobjects: Daily life among the Jupiter brains. *J. Evol. Technol.* **1999**, *5*, 1–34.
14. Aaronson, S. Guest column: NP-complete problems and physical reality. *ACM Sigact News* **2005**, *36*, 30–52. [CrossRef]
15. Shannon, C.E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]
16. Krauss, L.M.; Starkman, G.D. Universal limits on computation. *arXiv* **2004**.
17. Schaeffer, J.; Burch, N.; Bjornsson, Y.; Kishimoto, A.; Muller, M.; Lake, R.; Lu, P.; Sutphen, S. Checkers is Solved. *Science* **2007**, *317*, 1518–1522. [CrossRef] [PubMed]
18. Mahoney, M. Is There a Model for RSI? SL4. 20 June 2008. Available online: http://www.sl4.org/archive/0806/19028.html (accessed on 26 July 2018).
19. Turing, A. On computable numbers, with an application to the Entscheidungsproblem. *Proc. Lond. Math. Soc.* **1936**, *2*, 230–265.
20. Yampolskiy, R.V. Construction of an NP Problem with an Exponential Lower Bound. *arXiv* **2011**, arXiv:1111.0305.

21. Feuillet, L.; Dufour, H.; Pelletier, J. Brain of a white-collar worke. *Lancet* **2007**, *370*, 262. [CrossRef]
22. Johansson, B.B. Brain plasticity and stroke rehabilitation The Willis lecture. *Stroke* **2000**, *31*, 223–230. [CrossRef] [PubMed]
23. Bostrom, N. *Superintelligence: Paths, Dangers, Strategies*; Oxford University Press: Oxford, UK, 2014.
24. Kurzweil, R.; Schneider, M.L.; Schneider, M.L. *The Age of intelligent Machines*; MIT Press: Cambridge, MA, USA, 1990.
25. Smith, P. *An Introduction to Gödel's Theorems*; Cambridge University Press: Cambridge, UK, 2013.
26. Bringsjord, S.; Arkoudas, K. On the Provability, Veracity, and Al-Relevance of the Church—Turing Thesis. In *Church's Thesis after 70 Years*; Ontos Gmbh: Leipzig, Germany, 2006; Volume 1, p. 66.
27. Sotala, K. Advantages of artificial intelligences, uploads, and digital minds. *Int. J. Mach. Conscious.* **2012**, *4*, 275–291. [CrossRef]
28. Muehlhauser, L.; Salamon, A. *Intelligence Explosion: Evidence and Import, Singularity Hypotheses*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 15–42.
29. Yudkowsky, E. *Levels of Organization in General Intelligence, Artificial General Intelligence*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 389–501.
30. Eden, A.H.; Moor, J.H.; Soraker, J.H.; Steinhart, E. *Singularity Hypotheses: A Scientific and Philosophical Assessment*; Springer Science & Business Media: New York, NY, USA, 2013.

MDPI
St. Alban-Anlage 66
4052 Basel
Switzerland
Tel. +41 61 683 77 34
Fax +41 61 302 89 18
www.mdpi.com

Information Editorial Office
E-mail: information@mdpi.com
www.mdpi.com/journal/information

MDPI
St. Alban-Anlage 66
4052 Basel
Switzerland

Tel: +41 61 683 77 34
Fax: +41 61 302 89 18

www.mdpi.com

ISBN 978-3-03936-484-8