proceedings

MDPI

Proceeding Paper

Hybrid Information Systems: Who Is in Control? †

Daniel Boyd

Independent Researcher, 1625 GN Hoorn, The Netherlands; daniel.boyd@live.nl; Tel.: +31-(0)6-504-311-08

† Presented at the Online Workshop Digital Humanism: How to Shape Digitalisation in the Age of Global Challenges? IS4SI Summit 2021, Online, 14–16 September 2021.

Abstract: Digital Humanism aims to prevent technologies from developing in ways that are detrimental to humanity. As the level of interaction between technologies and their users intensifies, it becomes inappropriate to consider the situation as one involving two separate entities. Instead, they together form an integrated Sociotechnical System. This article considers such systems from the perspective of Emergent Information Theory, which provides an explanatory framework for the origins and functions of both designed technological and evolved biological information-based systems. Based on this unified perspective, the systemic challenges of controlling systems that combine designed and self-organizing components are discussed. The conclusions drawn provide theoretical foundations to support the realization of the goals of the Vienna Manifesto.

Keywords: digital humanism; Sociotechnical Systems Design; Emergent Information Theory; Artificial Intelligence

check for updates

Citation: Boyd, D. Hybrid Information Systems: Who Is in Control? *Proceedings* **2022**, *81*, 109. https://doi.org/10.3390/proceedings2022081109

Academic Editors: Wolfgang Hofkirchner and Hans-Jörg Kreowski

Published: 6 April 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The stated goal of the Vienna Manifesto on Digital Humanism (https://dighum.ec.tuwien.ac.at/dighum-manifesto, accessed on 10 March 2022) is to "shape technologies in accordance with human values and needs". The Manifesto's reference to the "co-evolution of technology and humankind" acknowledges the importance of approaching the combination of the two as an integrated system. This insight is not new. In the post-WWII years Sociotechnical Systems Design (STSD) was developed at the Tavistock Institute [1] to integrate the human and technological components of organizations to achieve their specified business goals. More recently, similar approaches have been applied to digital technologies [2]. When considering such digital sociotechnical systems, a unified theoretical framework is required for the informational aspects of the two components, such as that provided by Emergent Information Theory (EIT). This article combines STSD and EIT to address some of the specific challenges faced by Digital Humanism.

2. A Case Study

In 2016, Microsoft launched an experimental chatbot named Tay with the intention of developing "casual and playful conversation." While initially the bot produced pleasant responses, such as "*can I just say that im stoked to meet u? humans are super cool*", within 16 h it was posting material such as "*I fucking hate feminists and they should all burn in hell*". Tay was immediately shut down. The problem? Tay's lack of appreciation of contextual meaning made it vulnerable to exploitation by a specific group of people [3]. Microsoft's follow-up chatbot, Zo, successfully avoided falling into the same trap, but only by being strictly programmed to not respond to any term that could be an indicator of a controversial topic. This avoided Tay's excesses, but the strict limitations prevented Zo from fulfilling its intended function as an engaging chatbot for well-meaning human counterparts, a failure that eventually also led to shut down [4].

In neither case was the intended outcome achieved. The attempt to "shape technologies in accordance with human values and needs" failed in exactly the way that Digital

Humanism aims to prevent, and this example is anything but unique. The question is how this failure could have been averted.

3. Design vs. Self-Organization

In its reference to the "co-evolution of technology and humankind" the Manifesto uses the term 'evolution' in a generic sense to encompass any kind of progressive development. There are two ways in which complex functional systems can come to exist: design and self-organization. Design explains the existence of all physical technologies, computer hardware, and programmed software. Self-organization explains the existence of all natural and living systems, including the brain [5]. It can also underlie the emergence of function in Artificial Intelligence systems that have been so designed [6].

Functional designs are produced by an external designer to achieve a predefined purpose and then used to construct the solution from the necessary components. Both mechanism and outcome are known in advance. Self-organization refers to systems in which emergent phenomena arise spontaneously through interaction between the components of the system and often between the system and its environment [7]. The outcome in such situations is a result of the internal dynamics of the system itself and interaction with the environment and consequently, in all but the simplest examples, cannot be predicted in advance. The only way to discover the outcome is to wait for it to arise or run a simulation, which is essentially the same thing. The mechanisms by which the outcome is achieved are often complex and not easy to understand.

Functional design and self-organization each have their own strengths and weaknesses. Designed systems, due to the fact that they are constructed by external forces, lack the ability to maintain themselves and will inevitably fail unless actively maintained and repaired by these external forces. Their fixed input-output relationships also make them functionally 'brittle'; they cannot adapt to changing requirements. However, it is relatively easy to create a system with the required functionality, and improvements can be directly realized by changing the design.

In contrast, self-organizing systems have the advantage of self-correcting on perturbation and are adaptable to changing external and internal conditions. This is an advantage as long as it leads to the desired outcome; if not, then a very different kind of intervention is required than with designed systems [8]. Rather than directly correcting the outcome through redesign, the self-organizational processes can only be nudged in the desired direction as they proceed, exploiting their sensitivity to external conditions. For this to work, the change must not be in opposition with the system's internal tendencies. Otherwise, it will require considerable force and, as soon as the constraint is removed, the system will revert to its inherent state.

4. Physical vs. Informational Systems

The Manifesto's explicit focus on digital technologies emphasizes the distinction between physical and informational systems and functions. On the human side, we see the evolution of physical characteristics through the course of biological evolution taking place over millions of years. In technologies with physical functions, we also see functional progression, such as in the efficiency and power of the internal combustion engine. Alongside these physical functions, we see the evolution of the human brain and of digital technologies as physical systems supporting informational functions. However, these developments are no more than prerequisites; it is what goes on inside the brain and digital technologies that is of interest to Digital Humanism.

In both cases, we can observe physical input (via sense organs and digital input devices) leading to physical output (via expressed behavior and video output), but the processes that determine the input-output relationship are informational, not physical. Understanding these systems and the interactions between them that underlie their co-evolution therefore necessitates a unified approach to this aspect of their nature and operation, such as that provided by Emergent Information Theory.

5. Emergent Information Theory

Emergent Information Theory deals with the unique and very specific type of 'information' that underlies the function of designed and evolved information systems, such as computers and brains [9]. This differs from the more common definition of information that relates to the specification of a physical system, where the information and the system described are separate ("the map is not the territory"). A cornerstone of the theory is the observed dissociation between the digital information stored and processed in computers and the physical substrate this information is correlated with. Specifically, in contrast with the physical properties of an isolated magnetic bit, it is not possible to directly detect the binary information associated with it. Instead, this can only be deduced from the design of the system and the cause of the state. If the system were designed in a different way, the same information could be associated with either the opposite magnetic state or something entirely different, such as an electrical pulse, flash of light, or burst of sound. The theory therefore concludes that, while such information is entirely dependent on its physical substrate, it is inappropriate to consider it as a property of this substrate. Instead, it deserves to be considered as a non-physical substance in its own right.

This dissociation becomes more apparent when we consider the types of information we require of computers. These are not the binary values associated with individual bits but information, such as numbers, texts, and images, with an entirely different nature and properties. Creating such information requires the combination of large numbers of binary values into stacked hierarchical emergent levels, each with properties that cannot exist in its component parts: 8-bit bytes have decimal values that cannot be associated with an individual bit and RGB values cannot be associated with any one of the three bytes they are composed of. Significantly, these hierarchical levels are not created by interactions between the properties of the physical substrate but by logical relations between the information they carry. The same applies to the functions required of computers. The individual bit has no useful function. The creation of software involves the construction of hierarchical layers (machine-code operation, code line, subroutine, module), each with irreducible properties that are not present in its components.

Being non-physical, these emergent entities and functions cannot be directly detected or influenced using physical instruments. They can only be constructed by altering the physical entities (bits) that are associated with their fundamental components (binary values) according to strict coding systems. Extraction of the output is only possible by applying the same coding system to the measurable bits after the information process has run.

The reason why acknowledgement of this dissociation is important lies in the fact that, through the way we design computers, we grant these emergent non-physical functions top-down causative power over the events taking place in the physical substrate. The input-output relationship in computers is therefore not determined by physical laws, but by the high-level informational laws we design and build into them.

This hierarchical functional design does not stop at the level of the individual computer. Just as many machines with different physical functions are required to build a car, many computers with different informational functions are often required to create systems with more complex functions [10]. These computers are physically connected, but the function of these connections is not to create some combined physical effect (as in the machines in the factory) but allow their informational functions to be integrated into a higher-level emergent system.

What does this tell us about the second participant in Digital Humanism: the human brain? Extreme caution is required when comparing the two. Firstly, biological systems are orders of magnitude more complex than any technology. This is evident in comparisons of biological neural networks with the artificial networks that are frequently used to model them [11]. The physical mechanisms used by computers and the brain are also radically different, as are their morphological architectures. Even more significant are their fundamentally different ontologies: brains are autonomous, evolved, self-organizing systems while computers are dependent, designed, constructed machines.

In spite of these differences, there are also significant commonalities.

- Computers differ from other machines in having informational rather than physical functions, and brains differ from other organs in the same way.
- Given an isolated physical neural property, it is impossible to know what information is associated with it without knowing its cause and context. This is similar to the dissociation seen between bits and binary values.
- The information and functions associated with these basic components is not of the type required of the system, and the combination of these components to create these higher-level emergent functions is based not on their physical properties, but on the relationships and interactions between the information associated with them.
- While physical machines and organs are designed or evolved to perform a specific function, digital technologies and brains display extreme multifunctionality. From Turing's first conception of the programmable computer, it was clear that the same physical device could in theory perform any algorithmic function. Similarly, an immense diversity of informational processes can be supported by the same physical human brain.
- Associated with this functional flexibility is the fact that the development of new functions requires no changes in physical mechanisms. In this, both differ from physical machines and organs. Developing a novel software function does not require a different computer, and progressive intellectual progress has allowed the development of capabilities radically different to those of our prehistoric predecessors using physically identical brains.
- As with computers, intensive information exchange between individuals can lead to the creation of higher level (social) systems with properties that cannot exist at the level of the individual.

These commonalities between computers and brains allow their functions to be combined though information exchange between them, creating the integrated Sociotechnical Systems that are of concern to Digital Humanism. However, before considering how the function of these systems can be "shape[d] ... in accordance with human values and needs", we first need to consider the origins of the functions of each component.

6. Origins of Human Informational Functions

6.1. Self-Developed Functions

In the course of a human life, many informational functions are independently developed through self-organization of the brain's neural networks on the basis of internal mental processes and interaction with the outside world [12].

6.2. Taught Functions

If each human had to re-discover every wheel, we would never have progressed beyond our primeval origins. The progress of humanity is based on the transmission of skills and knowledge between and within generations. Examples are scientific theories and manual skills. Such mental abilities and ensuing behaviors are rationally designed by an expert to achieve some predefined purpose and then transferred through directed training. The design is subsequently improved through cyclical evaluation of performance, the directed solution of identified problems, and the introduction of innovations. This process results in stepwise improvements towards a desired input-output relationship. The same can be seen in social systems, such as legislative and political systems, which dictate the social behaviors and interactions of individuals to achieve societal purposes [13], and in the organizational optimizations of Sociotechnical Systems Design [1].

While the input-output relationships exhibited by individuals in these situations may be predetermined, this is not to say that the mechanism underlying the process is designed. On the contrary, the detailed neural pathways in different brains supporting identical behavior may be very different. Such functions are therefore created through a combination of design and self-organization.

Another significant feature of such designed functions is the dependence on compliance, which is unreliable for a number of reasons [14]. Frequently, there will be more than one design that the individual could follow, necessitating a choice. More significantly, these designs impinge on the existing intellectual and emotional environment of an individual with personal motivations and morals. The individual may disagree with the purpose of the design, or with the design as the means of achieving the purpose.

6.3. Self-Design vs. Self-Organization

One of the contributions of Sociotechnical Systems Design (STSD) was to acknowledge and address the limitations of external design. The initial successes in coal mines were the result of delegation of control to the operational self-organizing team in recognition of the fact that people doing a job may be more able to produce the best designs rather than some distant manager [15]. Here, the term 'self-organization' is not used in the systemic sense, since the outcome is still a designed solution constructed and maintained with the intention of creating a fixed effective relationship between input and output. Later developments in STSD took a more literal form of self-organization at the group level to reap the benefits of inherent adaptability to changing or unpredictable circumstances and robustness to perturbation [16]. Self-organizational processes also underlie the more generic development of informal social structures [17].

7. Origins of Digital Technological Functions

7.1. Programmed Software

The easiest form of control is seen in programmed software, which uses designed mechanisms to generate a specific output on the basis of a specific input [18].

7.2. Trained Artificial Neural Networks (ANNs)

Reliable input-output relationships (for instance pattern recognition tasks) can also be guaranteed through various forms of rigorous directed training in ANNs [19]. While the output may subsequently be predictable, the underlying mechanisms are often obscure, particularly in large-scale deep neural networks. 'Correction' of undesirable output can therefore only be realized through more training, not by adaptation of a design.

7.3. Self-Organizing Artificial Intelligence

Major benefits are to be gained from AI systems whose output is not predetermined by design but develops through use and interaction with the problem space. Firstly, this reduces the effort required to create the solution; secondly it can lead to solutions that could not have been created by design; and thirdly such systems are adaptable to unpredictable changes in the circumstances and problem to be solved. In this way they take a step away from specifically tooled problem solving towards general intelligence. This added value is inevitably coupled to reduced transparency and controllability [20]. The specific solutions they develop are not directly determined by human design. However, such self-determining systems still require humans to design and construct them, to present them with a problem, and to evaluate their output. In this way, humans can always regain control if the system develops in an undesirable direction, with discontinuation as ultimate sanction, as seen in the case of Microsoft Tay.

8. Discussion: Hybrid Information Systems

When designed and self-organizing components are combined to create an integrated system, the input-output relationship becomes more difficult to predict and influence. The case study of the Microsoft chatbots Tay and Zo is one example of such mixed, hybrid systems. One component was a digital system that was not designed to produce a fixed input-output relationship but to adapt to its interactions with humans. The other component was a group of people for whom no design had been made (only an assumption of moral behavior).

The outcome was the emergent product of the interaction between these two functionally adaptable components. When this went wrong, only the chatbot could be directly altered, but since it was not the source of the problem, this did not result in its resolution.

Another significant feature of this example is the lack of agency in the technological component, which has no motivation to achieve a particular goal and is fundamentally amoral. The human component, in contrast, does have agency and can introduce morality and immorality into the system: a distinction that makes it essential to avoid succumbing to the temptation to consider humans as technology or technologies as human [21].

To solve this problem, the creators of Zo reverted to design, programming explicit no-go areas. However, since the technology itself has no inherent value system, such designed solutions will be inevitably and undesirably brittle. This is exactly what was seen in the result, which prevented innocent discussion of certain topics while still failing to block all undesirable results.

A second tendency that needs to be resisted is that of approaching humans and computers as physical systems. Emergent Information Theory demonstrates just how detached the processes relevant to these challenges are from their physical substrates. It is only by acknowledging the informational nature of these phenomena that they can be sufficiently well understood to influence and control them.

9. Conclusions

This article considers Digital Humanism's ambition to "shape the co-evolution of technology and humankind in accordance with human values" in the context of Sociotechnical Systems Design (STSD) and Emergent Information Theory (EIT). It concludes that this challenge necessitates understanding of the nature of the non-physical integrated systems that form through the interaction between the informational processes supported by the human brain and digital technologies.

Consideration of the combination of designed and self-organizing components within an integrated system leads to the conclusion that creating a desired outcome of the co-evolution of technology and humankind cannot be realized simply by altering the design of the technological component, as demonstrated by the case study of the Microsoft chatbots Tay and Zo. Very different systemic approaches will need to be developed.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: I would like to acknowledge the interest, support, and constructive criticism of Wolfgang Hofkirchner, without which the relationships between Digital Humanism, Emergent Information Theory, and Sociotechnical Systems Design would not have become apparent to form the foundation of the content of this paper.

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

References

1. Trist, E.; Murray, H.; Trist, B.; Emery, F. *The Social Engagement of Social Science, Volume 2: A Tavistock Anthology—The Socio-Technical Perspective*; University of Pennsylvania Press: Philadelphia, PA, USA, 1990.
2. Winby, S.; Mohrman, S. Digital sociotechnical system design. *J. Appl. Behav. Sci.* **2018**, *54*, 399–423. [CrossRef]
3. Schwartz, O. In 2016, Microsoft's Racist Chatbot Revealed the Dangers of Online Conversation. *IEEE Spectrum*, 25 November 2019. Available online: https://spectrum.ieee.org/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation (accessed on 10 March 2022).
4. Stuart-Ulin, C. Microsoft's Politically Correct Chatbot Is even Worse than Its Racist One. *Quartz Ideas*, 31 July 2018. Available online: https://qz.com/1340990/microsofts-politically-correct-chat-bot-is-even-worse-than-its-racist-one (accessed on 10 March 2022).

5. Woese, C. A new biology for a new century. *Microbiol. Mol. Biol. Rev.* **2004**, *68*, 173–186. [CrossRef] [PubMed]
6. Prokopenko, M. Design versus self-organization. In *Advances in Applied Self-Organizing Systems*; Springer: London, UK, 2013; pp. 3–21.
7. Whitesides, G.; Grzybowski, B. Self-assembly at all scales. *Science* **2002**, *295*, 2418–2421. [CrossRef] [PubMed]
8. Gershenson, C. *Design and Control of Self-Organizing Systems*; CopIt ArXives: Mexico City, Mexico, 2007.
9. Boyd, D. Design and self-assembly of information systems. *Interdiscip. Sci. Rev.* **2020**, *45*, 71–94. [CrossRef]
10. Lock, R.; Sommerville, I. Modelling and analysis of socio-technical system of systems. In Proceedings of the 15th IEEE International Conference on Engineering of Complex Computer Systems, Oxford, UK, 22–26 March 2010; pp. 224–232.
11. Beniaguev, D.; Segev, I.; London, M. Single cortical neurons as deep artificial neural networks. *Neuron* **2021**, *109*, 2727–2739. [CrossRef] [PubMed]
12. Johnson, M. Functional brain development in humans. *Nat. Rev. Neurosci.* **2001**, *2*, 475–483. [CrossRef] [PubMed]
13. Koß, M. *Parliaments in Time: The Evolution of Legislative Democracy in Western Europe, 1866–2015*; Oxford University Press: Oxford, UK, 2018.
14. Johnson, R.; Chang, C.; Yang, L. Commitment and motivation at work: The relevance of employee identity and regulatory focus. *Acad. Manag. Rev.* **2010**, *35*, 226–245.
15. Trist, E.; Higgin, G.; Murray, H.; Pollock, A. *Organizational Choice: Capabilities of Groups at the Coal Face under Changing Technologies*; Routledge: London, UK, 1963.
16. Naikar, N.; Elix, B. Designing for self-organisation in sociotechnical systems: Resilience engineering, cognitive work analysis, and the diagram of work organisation possibilities. *Cogn. Technol. Work* **2021**, *23*, 23–37. [CrossRef]
17. Hofkirchner, W. Self-organisation as the mechanism of development and evolution in social systems. In *Social Morphogenesis*; Springer: Dordrecht, The Netherlands, 2013; pp. 125–143.
18. Scott, M. *Programming Language Pragmatics*; Morgan Kaufmann: Waltham, MA, USA, 2000.
19. Da Silva, I.; Spatti, D.; Flauzino, R.; Liboni, L.; dos Reis Alves, S. Artificial neural network architectures and training processes. In *Artificial Neural Networks*; Springer: Cham, Switzerland, 2017; pp. 21–28.
20. Thórisson, K.; Nivel, E.; Sanz, R.; Wang, P. Approaches and assumptions of self-programming in achieving artificial general intelligence. *J. Artif. Gen. Intell.* **2012**, *3*, 1–10. [CrossRef]
21. Hofkirchner, W. Digital Humanism: Epistemological, Ontological and Praxiological Foundations. In *AI for Everyone? Critical Perspectives*; University of Westminster Press: London, UK, 2021; pp. 33–47.