AI

MDPI

Communication

AI and We in the Future in the Light of the Ouroboros Model: A Plea for Plurality

Knud Thomsen

Laboratory for Neutron Scattering and Imaging, Paul Scherrer Institute, 5232 Villigen, Switzerland; knud.thomsen@psi.ch

Abstract: Artificial Intelligence (AI) is set to play an ever more important role in our lives and societies. Here, some boundary conditions and possibilities for shaping and using AI as well as advantageously embedding it in daily life are sketched. On the basis of a recently proposed cognitive architecture that claims to deliver a general layout for both natural intelligence and general AI, a coarse but broad perspective is developed and an emphasis is put on AI ethics. A number of findings, requirements, and recommendations are derived that can transparently be traced to the hypothesized structure and the procedural operation of efficient cognitive agents according to the Ouroboros Model. Including all of the available and possibly relevant information for any action and respecting a "negative imperative" are the most important resulting recommendations. Self-consistency, continual monitoring, equitable considerations, accountability, flexibility, and pragmatic adaptations are highlighted as foundations and, at the same time, mandatory consequences for timely answers to the most relevant questions concerning the embedding of AI in society and ethical rules for this.

Keywords: Ouroboros Model; self-consistency; categorical imperative; negative imperative; accountability; tolerance

Citation: Thomsen, K. AI and We in the Future in the Light of the Ouroboros Model: A Plea for Plurality. *AI* **2022**, *3*, 778–788. https://doi.org/10.3390/ai3040046

Academic Editors: Pablo Rivas, Gissella Bejarano and Javier Orduz

Received: 19 July 2022
Accepted: 17 September 2022
Published: 22 September 2022

1. Introduction

There is a wide variety of views in the assessments of the potential threats and benefits of full General Artificial Intelligence (AGI). In short, artificial agents, e.g., robots and software agents, are about to reach a level of cognitive performance that has hitherto only been exhibited by human beings. It is now speculated that, in the next step, these artificial actors or creatures will irreversibly surpass humans.

For some, AGI represents some sort of dystrophy; at a singular point of no return, AI would trump human cognitive and general capabilities, take over the world, enslave everyone, and, in the extreme case, basically eradicate all human beings and their rich cultures. These are some of the warnings that have recently kindled discussions [1,2]. As the diametric opposite, one might put all hope in benevolent superhuman agents, perhaps even trusting that AI by itself would enable a better human life. Quasi-religious hopes have been raised that all problems will be solved, and some sort of eternal life has even been promised [3]. When employing the only directly available mold, i.e., human intelligence, a German proverb comes to mind: "Der Schelm ist, wie er denkt", which translates into something like "the rogue models others' intentions on his own inclinations". Worries as well as aspirations accordingly could be seen as mainly revealing the attitudes of their human proponents. In contrast with these extreme views, I argue here for a reasonable and cautiously optimistic assessment, as black and white hardly ever capture a full image. Realistically, AI will in no way end all human toil once and for all. As with other advances in technology, it will shift the contents, weights, and values of activities and their organization in societies. AI will aggravate previous problems, solve some old problems, and generate some new problems. Without precautions, AI will exacerbate existing inequalities.

As just one example, some AI techniques are actually being employed very beneficially in current everyday reality, such as for the analysis of X-ray images and the identification of COVID-19 [4]. Employing AI for the first time in times of crisis, such as for dealing with the COVID-19 pandemic, has in particular been considered ethically delicate and demanding [5]. "Accountability for Reasonableness" is a recently coined catchphrase describing health ethics that can be aptly generalized to a principle of validity [6]. Beyond the context of biomedical ethics, the basic values of beneficence, non-maleficence, autonomy, and justice matter; for engineers, very much the same should be demanded, such as adherence to a precautionary principle, i.e., avoiding unnecessary risks [7]. Technical knowledge can be accumulated and transferred without much loss, while the development of a good character takes decades and, in the case of humans, has (almost) to start from scratch each time. A worrying example is that it took the greatest minds to understand nuclear physics and render that understanding useful for power generation and bombs. This knowledge now cannot just be erased (unless humans are wiped off the planet). It could happen that much lesser minds have control over the use of very mighty means. A similar situation might happen with powerful AI techniques. A participatory approach with open and transparent communication between diverse stakeholder groups during the development of AI systems has been suggested to be the best way of tackling these challenges. In a recent global survey, global convergence on a handful of key ethical principles for AI (transparency, justice and fairness, non-maleficence, responsibility, and privacy) has been found [8].

It has previously been argued that a simple conceptualization of AI as a mere tool will not do justice to the principal novel qualities of agents empowered by AGI. Allowing some level of intelligence requires learning, which—even with special dedicated efforts—(almost) inevitably entails a reduction in external understanding, transparency, and controllability [1,5,9,10]. This is seen as a principled, general, and most likely severe problem, even if arguments can be made to occasionally limit transparency [10,11]. It will not boil down to a question of design or be constrained to an appropriate level of detail, as an intelligent and autonomous agent at some point will intelligently decide for itself what to disclose (or even to lie). Genuine and convincing explanations and justifications would be required at a minimum, in particular for the public acceptance of procedures and resulting decisions [11].

Known AI systems have expertise in rather limited domains, are brittle, and lack flexibility, e.g., in the sense of learning quickly [12]. It has been suggested that the current limitations of AI might be overcome by following the design principles that are exhibited by natural agents, such as those outlined by the biologically inspired cognitive architecture of the Ouroboros Model [13,14]. While progress has certainly been made, and principled arguments appear to be outdated, there still remains some way to go [15]. However, some questions informed by that perspective can be considered even now.

2. Ouroboros Model in a Nutshell

In this section, the Ouroboros Model is introduced briefly. A mechanistic process and a memory structure are proposed, which in their interplay are claimed to explain facets of the general operation of efficient minds. The foci of this cognitive architecture are consistency, problem detection, problem solving, and learning. Early implementations in safeguards and safety systems have demonstrated the basic functioning of the Ouroboros Model, albeit in very limited settings [13,16–18].

The very foundation for all action and cognition according to the Ouroboros Model is provided by memory records, i.e., hierarchically organized schemata, which can be comprised of everything that is accessibly presented (e.g., by neurons in a living brain). Accordingly, such representations can be combined and thus meta-representations as well as abstractions can be derived and in turn made available for use at a later date [19]. Schemata are laid down in a kind of active memory [13]. The basic sequence of processes harnessing this foundation are presented in Figure 1.

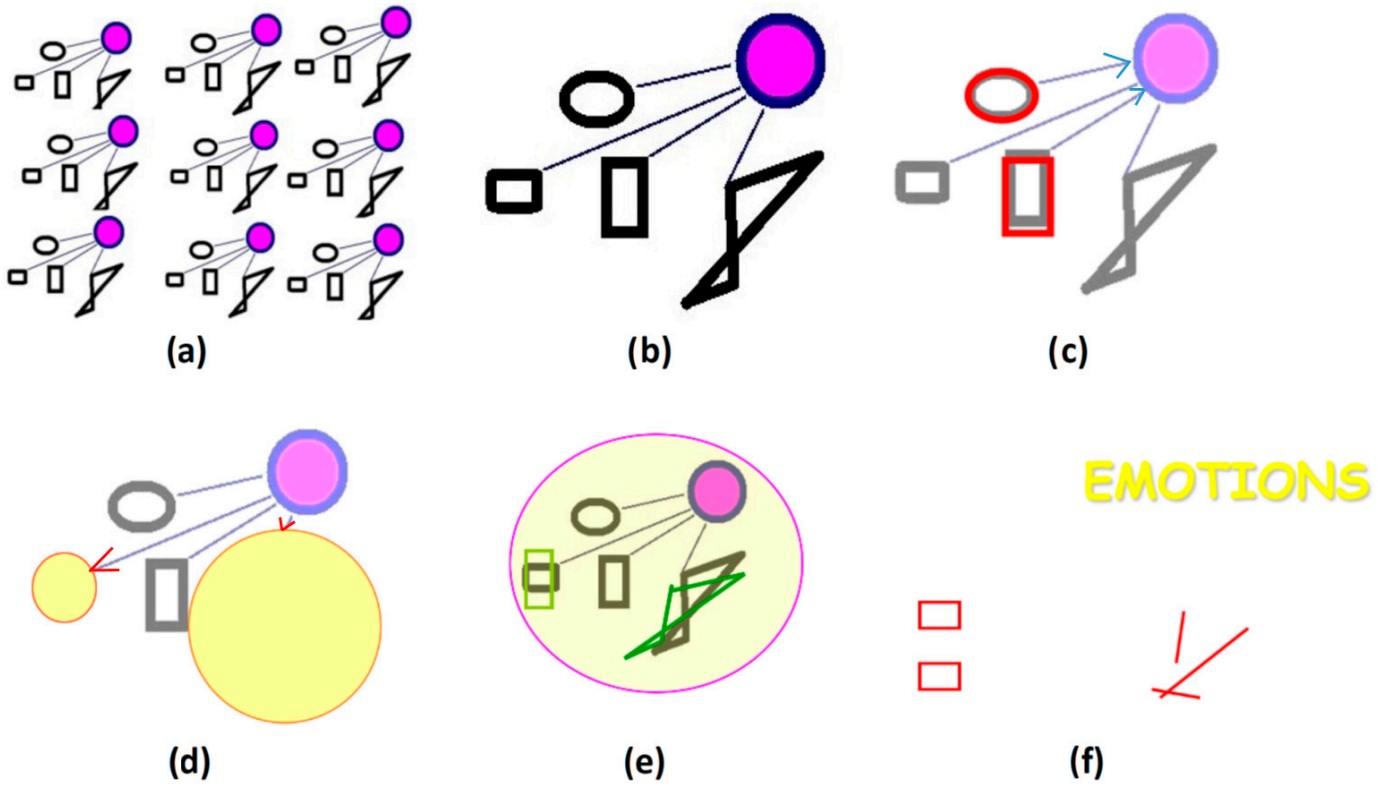

Figure 1. Basic structure and processes of the Ouroboros Model. (**a**) In a memory, there are many individual representations, i.e., (similar) schemata. (**b**) All schemata are comprised of specific combinations, i.e., linked bundles of features. (**c**) When an input becomes available, some initial features of a schema are activated. (**d**) As a consequence of the initial selection of a schema, the remaining linked features that make up that (partially) activated schema are biased and effectively searched for via spreading activation. (**e**) At regular intervals, all activations are balanced; a monitoring function (consumption analysis) checks the extent to which anticipations have been met, i.e., whether expected features in a schema are actually activated by the current input, and, conversely, to what extent occurring activations can be properly assigned, i.e., "consumed" by an existing schema. (**f**) Discrepancies/inconsistencies between expectations and actual occurrences are highlighted, and a feedback signal for the goodness of fit is produced, e.g., a measure of the degree to which a question has been answered satisfactorily. On a short time scale, features unaccounted for are included in the next iteration. Over longer time scales (and weighted by the importance of the involved features in the prevailing context), a more general signal is generated that can be understood as the "feeling"-component of an emotion [13].

In the existing technical implementations of the Ouroboros Model, some inheritance from cybernetics and synergetics is clearly visible; templates and thresholds are compared to actual measurements in order to assess a situation or detect some type of deviation or malfunction [16–18]. The used references themselves are obtained, i.e., the system is set up and calibrated, on the basis of earlier (successful) observations.

The Ouroboros Model stands in the tradition of perception–action cycle models and Bayesian inference, putting emphasis on distinct stages of the internal processing of an actor [20,21]. Its main innovation lies in proposing a clearly structured memory base and making manifold use of predictions and self-referential (performance) monitoring. The sequential activation of schemata components can be understood as a generalization of simple IF –> THEN (production) rules; each constituent feature can activate an entire schema and subsequently bias its other linked attributes. The resulting production rules are essentially linked.

A simple example is a schema with five slots that have to be filled for full activation of that particular schema. Any collection of four of these would be equivalent to "IF the one remaining slot can be filled THEN this schema is confirmed"; and, conversely, "IF this schema is applicable THEN this particular empty slot has to be filled in a specific way". The latter schema highlights how biasing works for features that have been found to be lacking. This is prominently harnessed for search applications. In the case that three or fewer of the five features are initially activated, the same applies but to a lesser extent; a distribution of the bias over, e.g., two open slots results in a reduced activation from that schema and a wider distributed search. In a setup where schemata compete for activation, reduced confidence in one assignment/interpretation can allow for possible alternatives to gain strength. Embedding nested loops in a (partially) hierarchical manner yields almost an unbounded degree of flexibility for the combination of schemata and the interpretation of input features. The same applies for actions.

With its basic processes including IF –> THEN rules as a kind of special case, it can be safely assumed that many of the results achieved with production systems are replicable by employing the Ouroboros Model. Full and partial activation of features in a given schema can be construed as implementing a version of Bayesian inference [13].

"Consumption analysis" is a recursive (self-)monitoring process that autonomously checks for the fulfillment of anticipations, directs attention, performs actions, and generates new schemata and stores them in relevant situations for later use where and when the need arises [19]. The process is the same regardless of the involved features, and it extends from simple perceive –> act schemata, which are required to perform specific routine tasks, to concepts for high-level self-reflection and goal setting. On the next level, the same basic processes apply during interactions between actors [22]. Following a continual incremental process, the schematic structures ever expand in a self-steered autocatalytic manner [19]. When thus grounding cognition in a memory structure, which itself is only built up as a result of an ongoing performance, no vicious circle arises; already-existing memory records (schemata) are leveraged and only adapted and further developed for later iterations in a subsequent step. In the case of living beings, grounding goes all the way down to the body of an actor. It can be hypothesized that something similar would be required for artificial agents if they need to exhibit some type of (self-)consciousness; this as well as "common sense" would then be somewhat "colored" by the basic setup of a particular implementation [22,23].

Importantly for transparency and an understanding of the developments from the outside, not much detail of this pivotal semantic content can be anticipated much earlier than it actually builds up. Some remedies for bringing light into the black boxes of AI, especially deep neural networks (DNNs), have been proposed. Relations to known models and concepts have been found to be decisive for making their internal operation somewhat understandable, and, at the same time, the basis laid down by explainable concepts (schemata) benefits the generation of suitable and efficient DNNs [14,23,24].

In a communication scheme, gaps in understanding/knowledge (in schemata) can often be filled by asking someone [22]. From the perspective of the Ouroboros Model, questions are prime examples of where a missing part of a schema can be (partially) constrained; questions about Who/What/When/Where/Why very clearly point to specific missing information. Scientific research can vary from fully exploratory to confirmatory and focused on scrutinizing previous results. Rigorously constraining questions and procedures greatly helps to render research understandable and reproducible [25].

At any given point in time, the cognitive power of an agent according to the Ouroboros Model is based on and limited by the availability of finely differentiated schemata (also known as mental models or narratives) that adequately cover relevant content in the widest and most consistent manner possible [13,14].

Special emphasis is placed on the "compartmentalization" of the (human) memory into partially disjunct schemata, which are organized in a non-strict hierarchy. Distinctions first start with dichotomies. During the process of development and learning, more

and finer details are added, and many simple dichotomies will be overturned at some point [26,27]. For humans, schemata are built up over a long period of growing, learning, and education [19,22].

The effective segmentation of memory underlies efficient cognition by employing (self-)consistency in various ways [28]. The linking of features into schemata defines what belongs together in a given context/situation, and thus it limits the number, extent, and weight of features to be considered. Still, when moving up the hierarchy, the quantity of lower-level constituents increases. Compartmentalization enables an efficient search for an anticipated but absent input, and we can actually draw inferences from the absence of an expected feature [29]. In a broader perspective, the tessellation of the accessible feature space by records facilitates simple amendments and confers some flexibility, such as allowing for the (interim) fixation of conflicting (disjoint) memories. Most likely, in a more comprehensive frame, such contradictions will surface, and they can be dealt with and possibly resolved at that higher level by taking additional information into account.

Generally, the combination of vastly different starting points and backgrounds, taking all of the resulting perspectives that are available at a particular moment into consideration, greatly enhances the potential for advantageous matches, responses, and promising adaptations.

3. Ethics in the Light of the Ouroboros Model as Deemed Relevant to AI

In this section, it is argued that rationally and self-consistently extending the functional model of cognition as outlined in the Ouroboros Model to contexts of human and human–artificial agent interaction will yield some sound guidelines without the need to resort to any grand metaphysics or ideology; bodily grounding and the evidence-based accumulation of mental structures do suffice.

What was sufficient in a specific situation is under optimum conditions clear after the fact. Encompassing and a thorough understanding, quite generally, are only approached in iterative cycles. For example, perhaps in a few years from now one will be able to say in hindsight which strategies for dealing with COVID-19 used by different states have proven to be the most appropriate (evaluations again will then depend on points of view, (ethical) priorities, etc.). Strict and immutable standards can only be set at the most abstract or general level (e.g., not endangering the survival of humankind including some positive perspective); very specific and unchangeable standards cannot do justice to dynamically evolving situations with an open future.

The Ouroboros Model has been claimed to, in a sense, embody Immanuel Kant's categorical imperative; nothing else but a general consistency condition is demanded, with no particular preferred individual [30]. As a downside, Kant excludes any qualification by the actual content and context. In perfect agreement with a compound view of interactions among humans, moral judgments have been found to be guided by a logic of universalization that combines and carefully balances rule- and outcome-based considerations [31]. Broadening that perspective to artificial agents is not straightforward for several reasons.

The biasing of open slots and entire schemata as described above does not only work on the level of simple, e.g., perceptual, schemata but also at the highest levels of abstraction. The basic goal of humans or, in general, living beings might be "survival" (and associated basic prescriptions/commandments as given by religions and philosophers). The grounding of the top-level aim in a body sets the stage for the overall architecture of natural actors. It is hypothesized that a similar drive for survival (continued existence) will predate the first flaring of consciousness in artificial agents.

Taking communication as an example, it is obvious that basic controversies often arise from differences in fundamental values and schemata, leading to biases at some intermediate cognitive levels [22]. Research has shown that such biases can arise and exist in many cases totally unconsciously [32]. In order to counteract these, and using only ethical arguments, several strategies have been proposed, some of which involve contextual information, the harnessing of external structures, and reliance on groups. Here, emphasis

is put on accountability, which has been found to be rather effective, probably to some extent due to the increase in the amount of attention that is paid to these issues [33].

At the same time as deriving high-level "Platonic" rules by abstraction, the Ouroboros Model emphasizes plurality, the all-importance of context, and striving for consistency. Except for the most strictly defined contexts, there is no guaranteed truth, no "absolutely right answer", and no unambiguous "opposite". To one or more of the relevant features of a schema a "NO-tag" can be assigned, while the other attributes should remain unchanged, which is not always possible and often not clear or easy to even identify. This becomes more complex as soon as time and evolution/growth (and erasure/forgetting) are taken into proper account.

The golden rule, especially in its positive formulation, has been declared to be strongly related to/rely on unreflective sympathy rather than true empathy [34]. It has to be "overcome" to some degree for a meaningful application across different individuals, diverse human backgrounds, and values, which possibly differ significantly between cultures in specific cases [34]. This appears to be a little easier in the case of a negative formulation of the golden rule. Attempting to apply the concept to artificial agents seems to stretch it (in particular, its positive formulation) much more, even beyond recognition for any detailed instructions concerning ethical action. A complex and equivocal relationship between morality and empathy has been found, and empathy has to be precisely defined when relating to humans; to comport with artificial agents, the concept has to be newly developed [35]. With a body and limits that stem from evolution over eons, taking the perspective of an electrical machine, for example, is not completely straightforward for any biological agent, and the same would apply the other way around as the implementations and resultant groundings are significantly different. This presupposes that comparable levels of cognition, reflection, self-awareness, and consciousness are actually accessible to both types of agents [1]. It has even been argued that (self-)consciousness will inevitably result from a certain level of cognitive performance [36].

A conscience as an internal reference at the top level of an individual is not so clear, nor is it identical for different humans. Here, it can only be hypothesized that it is based on some comprehensive and ideal self-schema and the evaluation of actions in that frame. The dichotomy of "good and bad" can easily be traced to the monitoring signal of consumption analysis; although there are always shades of grey, at some point a threshold is reached: a schema fits or it does not.

All rules, according to the Ouroboros Model, actually originate as advantageous abstractions from real, previously encountered, cases; they are "ground-in", abstracted, and established links in and between schemata when activated repeatedly. Whereas references to superhuman powers (gods) and the associated weights/biases might help believers adhere to rules and habits, atheists can at a minimum be just as consistent and moral [37]. In light of the above, it can even be expected that the (often) less-constrained views of disbelievers allow for greater varieties of meaningful options and an overall greater degree of rationality. In the absence of an extra-world "divine" justification, meaningful and transparent intra-world figures of merit are demanded, which have to be reflected upon, negotiated, and optimized, preferably in mutual agreement by all (potentially involved) parties.

According to the Ouroboros Model, only when all potentially relevant information is considered and taken into proper account can one hope to arrive at an optimum fit between problems (schemata with unfilled slots) and the solutions reachable at a given point in time [9,14]. On an abstract level, non-arbitrariness, fairness, and transparency can thus be identified as being part and parcel of this cognitive architecture and the involved hypothesized processes.

Following the outlined self-reflective iterative procedures, the weights of evidence and wishes have to be balanced. With fundamental limitations comprising many aspects, such as the total amount of existing resources, the lack of reserves, as well as the incomplete accessible experience base, insufficient memory capacity, sensitivity to discrepancies, speed of processing, and time available, this simply means that simplifications, compromises,

and risks are unavoidable; no interesting issue can be ever discussed exhaustively until it is resolved. Even so, (moral) nihilism at its core is self-contradictory [38]. Considering many individual and diverse points of view based on diverse histories and their different associated weights is the best method one can employ at any given time; tolerance for dissenting views follows naturally.

"Taking everything into account" does not mean "infinite" links nor any sort of indiscriminate associationism as relevancy is guaranteed by meaningful and bottom-up grounded compartmentalization. Actors should then be accountable in the sense that they have done their best to acquire, make available, and use all potentially relevant information at any given point in time. There certainly are practical limits to the cognitive capacities of actors and, even on a most basic level, not all situations allow for clear and defendable decisions; the unintended lurks almost everywhere [22]. A lack of time is but one constricting boundary condition for any action. Even for problems where the best solution would be available in principle, it is not at all clear that they could be found in a timely manner, nor that the concerned humans can agree on them. Still, not trying to identify the best and striving for largely acceptable solutions means wasting time.

4. Impact on and Further Recommendations to Societies

In a recent contribution, it has been proposed that looking at the extreme of the broadest available context leads to the conclusion of a valid "Negative Imperative" [9,38]. Irrespective of similarities in blueprints' details and implementations, on an overarching level, AI, natural beings, and states in this finite world are well advised to pay heed to that universal negative imperative. In a nutshell, given an inescapably limited overall frame containing only finite resources, we should avoid violence due to reflective self-interest [9,38].

In the real world with its ever-tighter constraints, acting is not a zero-sum game. Losses by one partner are not automatically a gain for another, and this has a more fundamental cause and goes much further than any simple retaliation. Most probably, any major damage caused by one individual means a loss for everyone, and in particular for the agent who was the initiator. Many others, who earlier were not at all directly involved, are then affected. A current example would be the war against Ukraine, with many people around the globe (who initially were completely uninvolved) being negatively affected.

In a sense, the negative imperative is similar to the golden rule, but it goes beyond the golden rule by setting forth a straightforward ("mechanistic") argument about why we should adhere to this directive. This means the founding of an "ought not" in rational egoism (or collective self-interest). It achieves the same for fairness as John Rawls' veil of ignorance, but in a much more natural, self-consistent, and evolutionary way while not suffering from any artificial or ideological limitations (except, perhaps, for demanding a certain level of prudence [38]). When emphasis is put on the overall boundary conditions and shared limitations, details concerning particular preferences lose their importance. Some qualified pacifism as anticipated by Immanuel Kant in his preference for republican/democratic structures (e.g., he thought that the very people who would suffer the most would never start wars) and later elaborated upon by Rawls with his theory of justice (containing the veil of ignorance) follows suit [39,40].

Humans should accordingly be much more afraid of natural stupidity—with its excessive craving for recognition or power, hubris, and malice—than of Artificial Intelligence, and this goes well beyond merely using grandiose names for limited functionalities [41,42].

Transparency and credible and consistent explanations have been emphasized as being most important for decision-making by humans as well as making artificial intelligence acceptable to those who are affected by it [5,10,23]. The inclusion of every individual who might possibly be involved is a well-established way of preventing (and later addressing) potential conflicts.

The above dovetails with what we now regard as falling under the title of good governance, i.e., public services being performed competently in the interests of all citizens

as outlined by the United Nations and elaborated upon a little more by the Council of Europe [43,44]. Recent findings on successful and prosperous human groups and states in history where the benefits of commonly produced goods and services are not conferred on a few, and citizens' wishes are taken care of, show that appropriate structures and processes predate modern liberal democracies [45]. Some well-organized institutions (i.e., a functioning bureaucracy) (have to) make sure that rules of law are adhered to, and systems of checks and balances in government have been developed by the founding fathers of many modern states to ensure that no one group or branch of society or government can become too powerful. In these prosperous societies, authorities have been found to be controlled and held accountable.

As with empathy, granting artificial agents "full democratic rights", as they are ideally assigned to humans, is not trivial. Any modulation of voting rights, e.g., taking expert knowledge into consideration when weighing actors' opinions and wishes, can very easily be rejected as discrimination.

Human experts often convene in groups and occasionally take turns when explaining their specific points and common findings, e.g., talking about COVID-19. Forecasts of the development of the pandemic as well as weather models, which are known to be sensitive to minute changes in starting (or boundary) conditions, are run in batches to determine how reliable the prognoses are. Something similar might be appropriate to implement in (AI) agents who make decisions or provide advice. In the real world, inequalities appear to be the norm rather than the exception (the "Pareto Principle"). Positive feedback is ubiquitous, especially in economic settings, which means that there exists a general ("natural") trend to enhance whatever inequalities arise. Counteracting these requires focused action. Quite generally, any expertise is inevitably be derived from the past, and it is often an open question as to how useful it will be in the unknown future. Experts, natural or artificial, bear the responsibility of explaining themselves, their background, and their reasoning and suggestions or decisions.

Similarly for the operation of a single brain, entire states, or AI, monopolizing power may possibly restrict and discourage valuable options, and, especially with changing requirements or boundary conditions, it might prevent the efficient identification of the best achievable solution for the whole. The principles of AI, compiled by the OECD and endorsed by G20 governments, provide a sound basis for development and progress [46,47]. Examples of issues in AI have been discussed, amongst others, by the World Economic Forum and include unemployment, inequality, humanity, stupidity, racism, security, unintended consequences, control, and robot rights [48]. In November 2021, after a lengthy consultation, a "Recommendation on the Ethics of Artificial Intelligence" containing some details was adopted by UNESCO's General Conference [49]. The "Ethics for Trustworthy AI" instrument, which takes into account input from more than 500 contributors, was devised by the European Union [50]. Attempting to pragmatically operationalize AI ethics is a next step [51].

For individual humans, whole societies, and mankind in general, AI allows us to bring more knowledge to bear for our common benefit than what would be possible without it. In particular, extensive data can be retrieved rapidly and exactly as available anywhere. Elaborate and knowledgeable consistency checks (necessarily AI supported) guard against forgeries and fake news. At present, but even more so in the long term, "thinking" machines, e.g., decision-support AI systems, will perform deep analyses of situations, and using them to recommend actions is certainly not only useful but inevitable. Transparent models and interactions are also key to facilitating the effective human use of (fallible) AI models [52]. Algorithms are far from perfect, and values can be endorsed or undermined. Accuracy, privacy, fairness and equality, property and ownership, accountability, and transparency "are sensitive to devaluation if algorithms are designed, implemented, or deployed inappropriately or without sufficient consideration for their value impacts, potentially resulting in problems including discrimination and constrained autonomy" [53].

Where it becomes especially tricky is sufficiently empowering AI systems to be able to stop (human or artificial) agents who are driven by nearsighted short-term (group-)egoist interests (i.e., who are focused on limited aspects taken into consideration from a very restricted perspective and understanding). Given changing boundary conditions, there is in general no fixed or stable optimum. Uninterrupted striving, which comprises continuous consideration and adaptation, is self-consistently required.

The resulting applicable figure of merit for any agent appears to be clear: what good he/she/it does for the whole and, in particular, for the weakest members of a community [7,38].

5. Conclusions

There is no (acceptable) way of turning back the clock or bringing it to a standstill: Artificial Intelligence will be further developed and, at some point, will continue to develop in a self-steered and self-paced manner with only very rudimentary outside control. Before that occurs, we need to establish a societal consensus on values, rules, and accountability for implementations along similar lines as proposed for human societies, in particular for ethical rules, which overarchingly could bind believers of different religions [7,54]. From a historical perspective, humans organized themselves in order to better support themselves, their kin, their tribe, or their state. At the current level of globalization and given the fact that we have already reached hard physical limitations, the frame or context to be taken into account can only be the entire world, and the appropriate level of reflection cannot be determined from any single point of view.

The Ouroboros Model, on a mechanistic, rational, and functional level, can motivate (if not justify) an overarching perspective that allows us to explain (on a coarse scale) cognition and its results, such as the rules and ethics that derive from the principled operation of an efficient and open mind. Some measure of pragmatism and tolerance in dynamic balance with basic accountability in this conceptualization result naturally from the inherent limitations of any agent at any given point in time.

Returning to the strongly contrasting attitudes described in the Introduction section, it might be advisable to develop and involve AI agents that can outwit limited selfish humans. These advisors ought to be "educated", take the "best" of human traditions (e.g., peaceful ones) as a basis, and consider relevant facts to the widest extent possible.

On the human side, capabilities entail obligations; what cannot be avoided needs to be shaped while "taking everything and everyone into balanced and reflected account". This ideal applies irrespective of the involved agents, whether they be human beings, states, artificial agents, or societies in general.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data are contained within the article.

Acknowledgments: Thoughtful and very constructive comments by the reviewers are gratefully acknowledged.

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

References

1. Meissner, G. Artificial intelligence: Consciousness and conscience. *AI Soc.* **2020**, *35*, 225–235. [CrossRef]
2. Musk, E.A.I. Doesn't Need to Hate Us to Destroy Us. *The New York Times*, 28 September 2020.
3. Kurzweil, R. *The Singularity Is Near: When Humans Transcend Biology*; Viking Books: New York, NY, USA, 2005.
4. van Ginneken, B. The Potential of Artificial Intelligence to Analyze Chest Radiographs for Signs of COVID-19 Pneumonia. *Radiology* **2021**, *299*, E214–E215. [CrossRef]
5. Cave, S.; Whittlestone, J.; Nyrup, R.; Eigeartaigh, S.O.; Calvo, R.A. Using AI ethically to tackle covid-19. *BMJ* **2021**, *372*, n364. [CrossRef]

6. Nyrup, R. From General Principles to Procedural Values: Responsible Digital Health Meets Public Health Ethics. *Front. Digit. Health* **2021**, *3*, 1–7. [CrossRef]
7. Hersch, M.A. Science, technology and values: Promoting ethics and social responsibility. *AI Soc.* **2014**, *29*, 167–183. [CrossRef]
8. Jobin, A.; Ienca, M.; Vayena, E. Artificial Intelligence: The global landscape of ethics guidelines. *Nat. Mach. Intell.* **2019**, *1*, 389–399. [CrossRef]
9. Thomsen, K. Ethics for Artificial Intelligence, Ethics for All. *Paladyn. J. Behav. Robot.* **2019**, *10*, 359–363. [CrossRef]
10. Héder, H. The epistemic opacity of autonomous systems and the ethical consequences. *AI Soc* **2020**. [CrossRef]
11. de Fine Licht, K.; de Fine Licht, J. Artifcial intelligence, transparency, and public decision-making. *AI Soc.* **2020**, *35*, 917–926. [CrossRef]
12. D'Amour, A.; Heller, K.; Moldovan, D.; Adlam, B.; Alipanahi, B.; Beutel, A.; Chen, C.; Deaton, J.; Eisenstein, J.; Matthew, D.; et al. Underspecification presents challenges for credibility in modern machine learning. *arXiv* **2020**, arXiv:2011.03395.
13. Thomsen, K. The Ouroboros Model in the light of venerable criteria. *Neurocomputing* **2010**, *74*, 121–128. [CrossRef]
14. Thomsen, K. The Ouroboros Model, Proposal for Self-Organizing General Cognition Substantiated. *AI* **2021**, *2*, 7. [CrossRef]
15. Stephan, K.D.; Klima, G. Artificial intelligence and its natural limits. *AI Soc.* **2021**, *36*, 9–18. [CrossRef]
16. Thomsen, K.; Pflügl, W.; Böck, H.; Hammer, J. In Proceedings of the Laser Surveillance System 7th Symposium on Safeguards and Nuclear Material Management, Liege, Belgium, 21–23 May 1985.
17. Thomsen, K. VIMOS, near-target beam diagnostics for MEGAPIE". *NIMA* **2007**, *575*, 347–352. [CrossRef]
18. Thomsen, K. Liquid metal leak detection for spallation neutron sources. *Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment.* **2008**, *592*, 476–482. [CrossRef]
19. Thomsen, K. Concept Formation in the Ouroboros Model. In Proceedings of the Third Conference on Artificial General Intelligence, Lugano, Switzerland, 5–8 March 2010.
20. Fuster, J.; Bressler, S. Past Makes Future: Role of pFC in Prediction. *J. Cogn. Neurosci.* **2015**, *27*, 639–654. [CrossRef]
21. Fröhlich, S.; Marcović, D.; Kiebel, S. Neural Sequence Models for Bayesian Online Inference. *Front. Artif. Intell.* **2021**, *4*, 1–17. [CrossRef]
22. Thomsen, K. The Ouroboros Model embraces its sensory-motoric foundations, Studies in Logic. *Gramm. Rhetor.* **2015**, *41*, 105–125. [CrossRef]
23. Linardatos, P.; Papastefanopoulos, V.; Kotsiantis, S. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy* **2021**, *23*, 18. [CrossRef]
24. Kazhdan, D.; Dimanov, B.; Jamnik, M.; Liò, P.; Weller, A. Now You See Me (CME): Concept-based Model Extraction. In Proceedings of the CEUR Workshop 2020 Proceedings, Virtual, 19–23 October 2020. [CrossRef]
25. Protzko, J.; Krosnick, J.; Nelson, L.D.; Nosek, B.A.; Axt, J.; Berent, M.; Buttrick, N.; DeBell, M.; Ebersole, C.R.; Lundmar, S.; et al. High Replicability of Newly-Discovered Social-behavioral Findings is Achievable. Available online: https://psyarxiv.com/n2a9x/ (accessed on 2 April 2022).
26. Rosenbloom, P.; (Computer Science, Univ. of Southern California, Los Angeles, CA, USA). private discussion via email, 2022.
27. Thomsen, K. It Is Time to Dissolve Old Dichotomies in Order to Grasp the Whole Picture of Cognition, in Theory and Practice of Natural ComputingFagan, D., Martín-Vide, C., O'Neill, M., Vega-Rodríguez, M.A., Eds.; TPNC 2018: Lecture Notes in Computer ScienceSpringer: Cham, Switzerland, 2018; Volume 11324. [CrossRef]
28. Thomsen, K. ONE Function for the Anterior Cingulate Cortex and General AI: Consistency Curation. *Med. Res. Arch.* **2018**, *6*, 1.
29. Dhurandhar, A.; Chen, P.Y.; Luss, R.; Tu, C.C.; Ting, P.; Shanmugam, K.; Das, P. Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives. In Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, QC, Canada, 3–8 December 2018.
30. Thomsen, K. What Immanuel Kant might have thought about the Ouroboros Model. In Proceedings of the The Fifth International Conference on Cognitive Science, Kaliningrad, Russia, 18–24 June 2012.
31. Levine, S.; Kleiman-Weiner, M.; Schulz, L.; Tenenbaum, J.; Cushman, F. The logic of universalization guides moral judgement. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 26158–26169. [CrossRef]
32. Correia, V. The Ethics of Argumentation. *Informal Log.* **2012**, *32*, 219–238. [CrossRef]
33. Correia, V. Contextual Debiasing and Critical Thinking: Reasons for Optimism. *Topoi* **2018**, *37*, 103–111. [CrossRef]
34. Bennett, M.J. Overcoming the Golden Rule: Sympathy and Empathy. *Ann. Int. Commun. Assoc.* **1979**, *3*, 407–422. [CrossRef]
35. Decety, J.; Cowell, J.M. Friends or foes: Is empathy necessary for moral behavior? *Perspect Psychol Sci.* **2014**, *9*, 525–537. [CrossRef]
36. Thomsen, K. Consciousness for the Ouroboros Model. *J. Mach. Conscious.* **2010**, *3*, 163–175. [CrossRef]
37. Ståhl, T. The amoral atheist? A cross-national examination of cultural, motivational, and cognitive antecedents of disbelief, and their implications for morality. *PLoS ONE* **2021**, *16*, e0246593. [CrossRef]
38. Thomsen, K. Gerecht und tolerant aus Vernunft und Eigeninteresse. *Aufklärung Und Krit.* **2017**, *92*, 82–109.
39. Kant, I. *Zum Ewigen Frieden: Ein Philosophischer Entwurf*; Reclam: Stuttgart, Germany, 2013.
40. Rawls, J. *A Theory of Justice, Cambridge, Massachusetts*; The Belknap Press of Harvard University Press: Cambridge, MA, USA, 1971.
41. McDermott, D. Artificial Intelligence meets natural stupidity. *ACM SIGART Bull.* **1976**, *57*, 4–9. [CrossRef]
42. Thomsen, K. Stupidity and the Ouroboros Model. In *Artificial General Intelligence, Lecture Notes in Computer Science*; Bach, J., Goertzel, B., Iklé, M., Eds.; Springer: Berlin/Heidelberg, Germany, 2012; Volume 7716, pp. 332–340.

43. Available online: https://www.ohchr.org/EN/Issues/Development/GoodGovernance/Pages/AboutGoodGovernance.aspx (accessed on 17 February 2022).
44. Available online: https://www.coe.int/en/web/good-governance/12-principles (accessed on 17 February 2022).
45. Blanton, R.E.; Fargher, L.F.; Feinman, G.M.; Kowalewski, S.A. The Fiscal Economy of Good Government: Past and Present. *Curr. Anthropol.* **2021**, *62*, 77–100. [CrossRef]
46. Recommendation of the (OECD) Council on Artificial Intelligence. Available online: http://mediaethics.ca/wp-content/uploads/2019/11/OECD-LEGAL-0449-en.pdf (accessed on 1 May 2021).
47. G20 Ministerial Statement on Trade and Digital Economy. June 2019. Available online: https://www.meti.go.jp/press/2019/06/20190610010/20190610010-1.pdf (accessed on 1 May 2021).
48. Bossmann, J. Top 9 Ethical Issues in Artificial Intelligence. Available online: https://www.weforum.org/agenda/2016/10/top-10-ethical-issues-in-artificial-intelligence/ (accessed on 17 February 2022).
49. Available online: https://en.unesco.org/artificial-intelligence/ethics#recommendation (accessed on 17 February 2022).
50. Available online: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (accessed on 17 February 2022).
51. Morley, J.; Elhalal, A.; Garcia, F.; Kinsey, L.; Mokander, J.; Floridi, L. Ethics as a service: A pragmatic operationalization of AI Ethics. *arXiv* **2021**, arXiv:2102.09364.
52. Nguyen, A.T.; Kharosekar, A.; Krishnan, S.; Krishnan, S.; Tate, E.; Wallace, B.C.; Lease, M. Believe it or not: Designing a Human-AI Partnership for Mixed-Initiative Fact-Checking. In Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology (UIST '18), Berlin, Germany, 14 October 2018; pp. 189–199. [CrossRef]
53. Hayes, P.; van de Poel, I.; Steen, M. Algorithms and values in justice and security. *AI Soc.* **2020**, *35*, 533–555. [CrossRef]
54. Küng, H. *Projekt Weltethos*; Piper: München, Germany, 1991.