Communication **Open Access**

Knud Thomsen*

Ethics for Artificial Intelligence, Ethics for All

https://doi.org/10.1515/pjbr-2019-0029
Received March 26, 2019; accepted September 12, 2019

Abstract: For human ethics, it can convincingly be argued that justice is a central cornerstone and basis. Here, it is suggested that this can, to some extent, similarly be applied to robots. The article makes the argument that Rawls' veil of ignorance in his conception of justice as fairness can effectively be replaced by a much more natural condition of prudent egoism in a finite world. Observing ones' own important interests in an encompassing context paves the way for a guideline for the conduct, which is binding for humans, robots and each and every pragmatic agent with a minimum level of rationality. These arguments do not see humans (forever) in any privileged position: any agent, single human, state, alien or artificial with a certain minimum of general cognitive (and effective) capabilities is bound by a universal negative imperative. This entails that precautious procedures are preferable, and some general prudently constrained flexibility is required for self-consistency and survival.

Keywords: social robots, ethics, justice, predictability, responsibility, Ouroboros Model, self-reflection, self-interest, negative imperative

1 Introduction

In this short opinion piece, an attempt is made to freshly sketch a very coarse but wide picture with the main aim of stimulating discussions on several diverse levels; a rather abstract conceptual layout is underpinned by concrete exemplary mechanisms.

The very concept of ethics is (hitherto) intimately linked to conscious and considerate human behavior. No baby, animal or simple inanimate object, including standard present-day robots, is expected to act ethically. Thus it is no wonder that whatever considerations might be undertaken, their starting point, or at least their inspirations, can be traced to concepts and rules formulated for adult and mentally healthy human actors. In the following, some observations, hypotheses, and arguments will be briefly introduced, all related to and from the perspective of a recent conception of a biologically inspired cognitive architecture, i.e., the Ouroboros Model. With motivations and guidance thus provided, this leads to results which themselves do not necessarily depend on the particulars of that model. A picture in very broad strokes will be presented, which is deemed necessary before delving into any detail in a next step.

1.1 Working hypothesis

It can be argued that for cases of similar abilities, both with respect to cognitive skills and actual physical or "impacting" abilities, artificial agents should be bound by a comparable ethic rules as natural agents (human beings).

1.1.1 This comes with an immediate first observation

Up until now, responsibility and accountability has been tightly tied to natural or legal persons, and, for the current state of affairs, there are good reasons to deny artificial agents, e.g., robots, legal personhood [1].

We see significantly different variants of behavior, all considered ethical by certain groups, and, whatever their exact content or form, existing rules and structures evidently do not work flawlessly for single humans and neither for human societies.

Still it can be argued that concepts developed for human interaction are the best available starting point for devising prescriptions for "ethical general action". It goes without saying that not each rule which has been developed in human societies in order to organize encounters and collaborations between their members, are relevant or applicable for human - robot interactions.

There seems to be plenty of existing material on regulating behavior to secure survival and improve human life and behavior by ensuring the adherence to general (ethic) rules. For humans, the Ten Commandments are an obvious example, and some of them also appear applicable to other agents more generally. For artificial agents in particular, Asimov's Three Laws of Robotics have been pro-

***Corresponding Author: Knud Thomsen:** Paul Scherrer Institut, CH-5232 Villigen PSI, Switzerland, E-mail: knud.thomsen@psi.ch

posed [2]. To these, many additions and amendments for specific cases, situations, and relevant circumstances and environments have been recommended [3–5], for instance:

- An artificial agent must be understandable
- All actions of an artificial agent must be transparent
- An artificial agent must be able to explain itself
- An artificial agent should always have an off switch

The prescriptions given above strongly presuppose a fundamental gulf between (conscious and free) humans and other, in particular, robotic, agents; – "Robots are multi-use tools…" is an example, setting the theme according to this mind frame in a recent statement by an expert group (Engineering and Physical Sciences Research Council) [6].

1.2 Intermediate thesis / postulate

Any distinction between robotic versus "virtual" agents (software) is of very limited importance; the only relevant issue for any action is whether a mechanism can produce consequences in the real world where other (human) actors thrive. In a dramatic self-consistent abbreviation: They should not (unduly) be made to suffer from any activity of that first agent.

Autonomy can easily be identified as the really decisive feature in this context, which is intimately tied to control and predictability, and, on higher levels: accountability, responsibility, liability, principles, rules, duties, purposes, intentions, plans, and goals, which are all tightly linked to each other and result from taking different perspectives and considering diverse grades of abstraction. These issues finally culminate in questions relating to free will and consciousness [7].

1.3 Second observation

Any "tooling argument", i.e., considering artificial agents strictly as simple tools for human purposes, falls short of the problem.

Nobody was ever concerned with the ethics of a stick, scissors, or diesel engines. Trying to address and solve issues of ethics and responsibility by definition means staying in a limited, narrowly contained frame; it does not work for advanced AI, and neither for robots:

It is exactly the real possibility for non-predictable, non-controllable, "self-steered" and "self-controlled", "autonomous" action by any artificial agent, which renders the topic interesting and relevant [7–9]. Actors sufficiently endowed with artificial intelligence, which are thus able to choose goals for themselves, constitute the important cases.

A well-known and relatively simple example is posed by neural networks. Powerful deep neural nets surpass human capabilities in several (so far, disconnected and limited) fields of expertise. Deep convolutional networks do not classify figures based on global shape, and often they don't use what humans would consider the relevant features [10, 11]. We often do not understand their working in any satisfactory detail; in many cases we have no clue why a certain outcome came about.

The Ouroboros Model can serve as another example. It is a biologically inspired cognitive architecture with an "algorithmic backbone" of "consistency curation" at its very core [12]. Monitoring and striving for general consistency is implemented by an iterative process, which works based on memories. These are organized into schemata, and, in turn, this substrate expands dynamically and autocatalytically as a consequence of the working process. The Ouroboros Model is briefly cited here because its set-up makes it very plausible that high levels of general sophistication cannot be achieved without equally expanding autonomy, and it can deliver inspiration and arguments for a universal guideline for conduct which any sufficiently intelligent agent has to follow [13–15].

2 Ouroboros model in a nutshell

The basis for cognition is seen as consisting in hierarchically organized schemata, which are laid down in a kind of active memory. A recursive self-monitoring process termed consumption analysis autonomously directs attention, action and also the generation of new schemata and their storage in relevant situations for later use where the need arises [12]. This extends from simple perceive –> act schemata required to independently perform specific tasks to concepts for high level self-reflection and goal-setting [7]. No details of this foundational content can be anticipated before it is actually generated and laid down in memory. In a dynamic incremental process, these schematic structures are ever-expanding; this particularly happens unpredictably, further enhancing the autonomy of the agent.

On another level, the same basic processes apply during interactions between actors. Reciprocity in a dialogue is a simple extension of the central roles of consistency and an adequate fit between the expectations and experience of the (communicative) interaction of several partners [16].

In step with the accruing memory basis, the cognitive capabilities of the acting agent increase. The very monitoring process of comparing expectations with actually encountered and available input stays the same; only the content changes, with the perception of simple attributes at the "lower" end of the scale and complex schemata, e.g., involving overarching goals and representations of the actor itself, "higher up".

The important point here is that nobody has full control over how the conceptual basis, i.e. the accumulating schemata of an agent implementing the Ouroboros Model, develops. In fact, the more the cognitive development and expansion progress, the more the number of options for new and flexible combinations increases and their predictability diminishes. The Ouroboros Model directs attention to the widest possible coverage and consistency of evidence, and, in particular, to the dimension of time. While cognitive structures can grow with few constraints, expansion affecting the real world finds its limits at hard physical boundaries.

3 Justice

For humans, "justice" is a most widely shared concept appreciated by vastly different cultures and at the core of ethical behavior, and, it seems, universally demanded as the basis for human interactions [14]. In an impressive philosophical undertaking in the last century, John Rawls has reasoned that justice is best understood as fairness, and his view has been widely acclaimed [17, 18]. Rawls devised a procedure for how to establish fairness in a democratic society by utilizing his "famous veil of ignorance", where decisions are taken and goods are distributed with everyone involved in complete ignorance of their own status. Each individual has to consider the possibility that she or he is the worst off, and therefore tries his or her best to further the not so lucky partners in the experiment [17, 18].

Alas, this argumentation runs completely counter to another central idea, indispensable for just decisions, in particular at a court. There, it is of utmost importance that publicly known and accepted rules are followed in a fully transparent manner, that decisions can be explained, and ideally they are justified by showing how they take into unbiased account all the applicable regulations and all available evidence as comprehensively and consistently as possible at and in the appropriate time.

The Ouroboros Model sees justice as an abstraction of repeated occasions where human actions affecting other humans give one the positive feeling of a well-balanced exchange, good relations, and fair (social) structures. At least with respect to the first point, this seems applicable to interactions involving artificial agents and robots. What is just demanded in hindsight, i.e. fully taking into account all relevant information, is strongly advised even before an action is begun.

3.1 What ought I to do?

Immanuel Kant posed that question long ago [19]. His concisely formulated answer is: "act so as if your maxims should serve at the same time as the universal law". This Categorical Imperative, in the light of the Ouroboros Model, is nothing but a general condition of consistency, with no particular preferred individual. This follows directly from Kant's demand that any action (at least its intention) should be suitable to serve as a general model and binding law for everyone.

Admitting a fundamental problem to obey Kant's prescription, in particular, with only limited time available, there is nothing like "Absolute Justice" compatible with the tenets of the Ouroboros Model. Decisive for humans is the shared intention of, first, survival and, subsequently, "improvement", i.e., striving for a common world with mutual respect, and a general appreciation of personal dignity, above all for actions compatible with widely shared wellbeing and sustainable progress [14].

3.2 "Negative Imperative"

With tight common restrictions and complex, interlaced links and dependencies between partners, any strongly violent action with high probability has negative impacts also on the originator and the whole world and thus should be avoided based on his or her intrinsic self-interest (or collective self-interest) [14]. This foundation in prudent egoism achieves basically the same for fairness as Rawls' veil of ignorance, but in a much more natural and fully self-consistent evolutionary way, and not bound by any artificial limitation (maybe except demanding a certain level of prudence [15]). Unavoidable ignorance or uncertainties concerning possible consequences strengthen this conclusion.

The decisive point, especially when comparing our present situation with previous times, is that not only do interactions and dependencies become more restrictive as the available frame quickly shrinks, but there is also no "reserve" or "buffer" left at an "outside". No additional resources from somewhere else can be brought to bear.

The accessible real world is all which can credibly be assumed, and every large change by any actor inescapably self-impacts the originator [14].

As the ultimate example of this, nuclear war knows no winner. When, in former times, armies fought against each other, no matter how strong and large, there was no danger to the survival of mankind. In contrast to this, nuclear winter does not invite for skiing and poses a severe threat to almost every living being. Something similar can be said about misusing the atmosphere and sea for dumping waste. At the scale of tribes strolling around the planet, waste disposal was no issue at all. At the current global scale, negative impacts affect everyone.

It is not "ought" which can be derived from "is" but "ought not", and a positively formulated corollary: try to be consistent in the best, i.e., widest accessible relevant frame.

This is not an unexpected result, and neither should it be seen as irrelevant. Looking on the Ten Commandments or, e.g., on modern-day signposts in any densely inhabited place, the rules forbidding something quite often outnumber positive prescriptions.

As the negative imperative can be reasoned for in a fully transparent and consistent manner, it can be seen as contradicting David Hume's famous finding that no "ought" can be derived from "is".

In fact, there is no contradiction; (only) under the assumption that an actor cares for something durable of value to her, strongly destructive actions are prohibited. Then, some obligations can in fact be derived from (the appreciated value of) something in existence.

3.3 The above does not see humans (forever) in any privileged position

Any actor, single human, state, alien or artificial, e.g., robot, with a certain minimum of general cognitive (and effective) capabilities, is bound by the negative imperative; precautious procedures adhered to over time and some prudently and pragmatically constrained flexibility ensue. This includes actions towards others, specifically human, actors, and means, in particular, refraining from any insistence on black and white (in particular, formal) rules without due wide-ranging considerations. An obligation to prudency can be claimed as a direct general consequence and self-consistently mandatory [14].

Likewise, whether any specific action is ethical or not does not primarily depend on a human being directly or immediately affected.

Following the general basic layout of processes in the Ouroboros Model based on iterative cycles and, in some sense inevitably resulting, incremental progress, the presented argumentation suggests a universal prescription for action. It is of no surprise and adds to the self-referential credibility that no specific commandments can be claimed as valid under all circumstances.

In a rather direct move, these conditions and guidelines for action can be applied self-reflectively to the topic in question. Returning thus to the second sense of "ethics for Artificial Intelligence" as could be read into the title of this brief paper, one could try to formulate interrelated rules which might be useful to observe when devising artificial agents.

4 How to build an intelligent, conscious and ethical artificial agent

Extending the line of arguments above for possible implementation in a self-reflective and self-consistent way, the Ouroboros Model gives the following very coarse draft sketch of seemingly essential ingredients for acceptable, ethical, and finally conscious, artificial systems including all agents and robots successfully sharing our real physical and virtual worlds:

- Large (compartmentalized) memory organized in hierarchies of rich differentiated and encompassing schemata
- Consistency Curation, i.e., discrepancy monitoring and its manifold and repeated use for flexibly steering them towards promising future actions

The above two bullet points, in particular, enable iterative self-reflective, self-controlled autocatalytic increments in the conceptual basis, where demanded by the context and circumstances, i.e., storage of new schemata

- An endowment with flexible, diverse and redundant means for sensing the environment and also for communication with the outside
- Learning and development depends on an assistive environment including, in particular, other agents; this necessarily also requires somewhat ordered and predictable conditions
- Time is essential for the growth of adapted useful schemata comprising representations of a self as well as for other agents

– Self-interest promoting curiosity, mutual respect and fairness, prudence, pragmatism, tolerance and modesty [14].

Admittedly, while these prescriptions can be argued for in the frame of the Ouroboros Model, they need not all necessarily be implemented when building working AI or, in particular, robots. Neither are all points of equal importance, especially not for the currently accessible levels of AI performance. Still, it is claimed that there are good reasons for following these steps and that, indeed, they are not arbitrarily chosen but self-consistently result from self-directed growth of cognition while, at the same time, promoting it.

5 Preliminary concluding remarks

Ethics for AI cannot be expected to be any simpler than ethics for humans. On the contrary, this has to be expected to be a much more complex and involved topic as at least one fundamental ingredient is certain to be different. Humans most likely will not so easily extend full compassion, which has been claimed to lie at the foundation of their ethical behavior amongst themselves, to artificial agents. Even less so will all artificial actors easily and "naturally" develop empathic feelings for humans and towards each other, although in the here proclaimed optimistic view growing rationality inevitably at some stage breeds (self-) consciousness, fairness and caution.

Matters are even more complicated as no clear-cut distinctions seem to persist. In hypothetical moral dilemmas, where humans attributed feelings to humanized robots, they were less likely to sacrifice them in order to save the lives of anonymous humans [9]. With social robots starting to be employed in therapeutic settings, psychological and ethical aspects become topics of pressing urgency [20].

In sum, one can argue for some general pragmatic rules which do not differ between specific types of agents and which do not depend on a prerequisite of good will but rather are founded on self-interested rationality, aiming at justice, tolerance and modesty in a finite and ever more constrained world. Agents thus should strive for establishing these as a good basis for ethical interactions which do not undermine a promising and free future [14].

Acknowledgments: Extraordinarily kind support from the editor and thought-provoking criticism by two anonymous reviewers are gratefully acknowledged.

References

[1] L. Floridi, M. Taddeo, Romans would have denied robots legal personhood, Nature, 2018, 557, 309
[2] I. Asimov, Runaround, I, Robot (hardcover) (The Isaac Asimov Collection ed.), New York City: Doubleday, ISBN 0-385-42304-7, 1950
[3] T. Hellström, S. Bensch, Understandable robots, Paladyn, Journal of Behavioral Robotics, 2018, 9, 110-123
[4] S. Wachter, B. Mittelstadt, L. Floridi, Transparent, explainable, and accountable AI for robotics, Science Robotics, 2017, 2(6), eaan6080
[5] D. Heaven, Robot laws: Why we need a code of conduct for AI – and fast, New Scientist, 4 August 2018, 38-41
[6] Principles of robotics - EPSRC website, retrieved 2018-08-22 https://epsrc.ukri.org/research/ourportfolio/themes/engineering/activities/principlesofrobotics/
[7] W. F. G. Haselager, Robotics, philosophy and the problems of autonomy, Pragmatics & Cognition, 2005, 13(3), 515-532
[8] W. Barfield, Liability for autonomous and artificially intelligent robots, Paladyn, Journal of Behavioral Robotics, 2018, 9, 193-203
[9] S. R. R. Nijssen, B. C. N. Müller, R. B. van Baaren, M. Paulus, Saving the robot or the human? Robots who feel deserve moral care, Social Cognition, 2019, 37(1), 41-52
[10] N. Baker, H. Lu, G. Erlikhman, P. J. Kellmann, Deep convolutional networks do not classify based on global object shape, Computational Biology, 2018, https://doi.org/10.1371/journal.pcbi.1006613
[11] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, K. R. Müller, Unmasking Clever Hans predictors and assessing what machines really learn, Nature Communications, 2019, Article number 1096, https://doi.org/10.1038/s41467-019-08987-4
[12] K. Thomsen, The Ouroboros Model in the light of venerable criteria. Neurocomputing 74, 2010, 121-128
[13] K. Thomsen, Concept formation in the Ouroboros Model, Third Conference on Artificial General Intelligence, Lugano, Switzerland, March 5-8, 2010
[14] K. Thomsen, Gerecht und tolerant aus Vernunft und Eigeninteresse, Aufklärung und Kritik 92, 2017, 82-109
[15] K. Thomsen, Stupidity and the Ouroboros Model, In: J. Bach, B. Goertzel, M. Iklé (Eds.), Artificial General Intelligence, Lecture Notes in Computer Science, Vol. 7716, (pp. 332-340), Berlin, Heidelberg, Springer, 2012
[16] K. Thomsen, The Ouroboros Model embraces its sensory-motoric foundations, Studies in Logic, Grammar and Rhetoric, 2015, 41, 105-125
[17] J. Rawls, Eine Theorie der Gerechtigkeit (translation Herrmann Vetter), Suhrkamp, Frankfurt am Main, 2014
[18] J. Rawls, Justice as Fairness, a Restatement, Harvard University Press, Cambridge, MA, London, England, 2003
[19] I. Kant, Gesammelte Schriften, Hrsg.: Bd. 1-22 Preussische Akademie der Wissenschaften, Bd. 23 Deutsche Akademie der Wissenschaften zu Berlin, ab Bd. 24 Akademie der Wissenschaften zu Göttingen, Berlin 1900 ff, AA IV, 421 / Weischedel 4, 51 / GMS 51-53
[20] E. Fosch-Villaronga, J. Albo-Canals, "I'll take care of you," said the robot, Paladyn, Journal of Behavioral Robotics, 2019, 10, 77-93