Error-Correction for AI Safety

Nadisha-Marie Aliman[1], Pieter Elands[2], Wolfgang Hürst[1], Leon Kester[2], Kristinn R. Thórisson[4], Peter Werkhoven[1,2], Roman Yampolskiy[3], and Soenke Ziesche[5]

[1] Utrecht University, Utrecht, Netherlands
[2] TNO Netherlands, The Hague, Netherlands
[3] University of Louisville, Louisville, USA
[4] Reykjavik University and Icel. Inst. for Intelligent Machines, Iceland
[5] Independent Researcher, Delhi, India

Abstract. The complex socio-technological debate underlying safety-critical and ethically relevant issues pertaining to AI development and deployment extends across heterogeneous research subfields and involves in part conflicting positions. In this context, it seems expedient to generate a minimalistic joint transdisciplinary basis disambiguating the references to specific subtypes of AI properties and risks for an *error-correction* in the transmission of ideas. In this paper, we introduce a high-level *transdisciplinary system clustering* of ethical distinction between antithetical clusters of *Type I* and *Type II* systems which extends a cybersecurity-oriented AI safety taxonomy with considerations from psychology. Moreover, we review relevant Type I AI risks, reflect upon possible epistemological origins of hypothetical Type II AI from a cognitive sciences perspective and discuss the related human moral perception. Strikingly, our nuanced transdisciplinary analysis yields the figurative formulation of the so-called *AI safety paradox* identifying AI control and value alignment as conjugate requirements in AI safety. Against this backdrop, we craft versatile multidisciplinary recommendations with ethical dimensions tailored to Type II AI safety. Overall, we suggest proactive and importantly *corrective* instead of prohibitive methods as common basis for both Type I and Type II AI safety.

Keywords: AI Safety Paradox, Error-Correction, AI Ethics

1 Motivation

In recent years, one could identify the emergence of seemingly antagonistic positions from different academic subfields with regard to research priorities for AI safety, AI ethics and AGI – many of which are grounded in differences of short-term versus long-term estimations associated with AI capabilities and risks [6]. However, given the high relevance of the joint underlying endeavor to contribute to a safe and ethical development and deployment of artificial systems, we suggest placing a mutual comprehension in the foreground which can start by making

references to assumed AI risks explicit. For this purpose, we employ and subsequently extend a cybersecurity-oriented risk taxonomy introduced by Yampolskiy [37] displayed in Figure 1. Taking this taxonomy as point of departure and modifying it while considering insights from psychology, an ethically relevant clustering of systems into *Type I* and *Type II* systems with a disparate set of properties and risk instantiations becomes explicitly expressible. Concerning the set of Type I systems of which present-day AIs represent a subset, we define it as representing *the complement of the set of Type II systems*. Conversely, we regard hypothetical Type II systems as *systems with a scientifically plausible ability to act independently, intentionally, deliberately and consciously*. Given the controversial ambiguities linked to these attributes, we clarify our idiosyncratic use with a working definition for which we do not claim any higher suitability in general, but which is particularly conceptualized for our line of argument. With Type II systems, we refer to systems having the ability to construct counterfactual hypotheses about what could happen, what could have happened and why including the ability to simulate "what *I* could do", "what *I* could have done" and the generation of "what if" questions. (Given this conjunction of abilities including the possibility of what-if deliberations with counterfactual depth about self and other, we assume that Type II systems would *not* represent philosophical zombies. A detailed account of this type of view is provided by Friston in [20] stating e.g. that *"the key difference between a conscious and non-conscious me is that the non-conscious me would not be able to formulate a "hard problem"; quite simply because I could not entertain a thought experiment*".)

How and When did AI become Dangerous		***External Causes***			***Internal Causes***
		On Purpose	***By Mistake***	***Environment***	***Independently***
Timing	***Pre-Deployment***	*a*	*c*	*e*	*g*
	Post-Deployment	*b*	*d*	*f*	*h*

Fig. 1. Taxonomy of pathways to dangerous AI. Adapted from [37].

2 Transdisciplinary System Clustering

As displayed in Figure 1, the different possible external and internal causes are further subdivided into time-related stages (pre-deployment and post-deployment) which are in practice however not necessarily easily clear-cut. Thereby, for Type I risks, we distinguish between the associated instantiations *Ia* to *If* in compliance with the *external causes*. For Type II risks, we analogously consider external causes (*IIa* to *IIf*) but in addition also *internal causes* which we subdivide into the novel subcategories "on purpose" and "by mistake". This assignment leads

to the risks *IIg* and *IIh* for the former as well as *IIi* and *IIj* for the latter subcategory respectively. The reason for augmenting the granularity of the taxonomy is that since Type II systems would be capable of intentionality, it is consequent to distinguish between internal causes of risks resulting from intentional actions of the system and risks stemming from its unintentional mistakes as parallel to the consideration of external human-caused risks a and b versus c and d in the matrix. (From the angle of moral psychology, failing to preemptively consider this subtle further distinction could reinforce human biases in the moral perception of Type II AI due to a fundamental reluctance to assign experience [25], fallibility and vulnerability to artificial systems which we briefly touch upon in Section 3.2.) Especially, given this modification, the risks *IIg* and *IIh* are not necessarily congruent with the original indices g and h, since our working definition was not a prerequisite for the attribute "independently" in the original taxonomy. The resulting system clustering is illustrated in Figure 2.

TYPE I CLUSTER

How and When did Type I system become Dangerous		External Causes		
		On Purpose	By Mistake	Environment
Timing	Pre-Deployment	*a*	*c*	*e*
	Post-Deployment	*b*	*d*	*f*

TYPE II CLUSTER

How and When did Type II system become Dangerous		External Causes			Internal Causes	
		On Purpose	By Mistake	Environment	On Purpose	By Mistake
Timing	Pre-Deployment	*a*	*c*	*e*	*g*	*i*
	Post-Deployment	*b*	*d*	*f*	*h*	*j*

Fig. 2. Transdisciplinary system clustering of ethical distinction with specified safety and security risks. Internal causes assignments require scientific plausibility (see text).

Note that this transdisciplinary clustering does *not* differentiate based on the specific architecture, substrate, intelligence level or set of algorithms associated with a system. We also do not inflict assumptions on whether this clustering is of hard or soft nature nor does it necessarily reflect the usual partition of narrow AI versus AGI systems. Certain present-day AGI projects might be aimed at Type I systems and some conversely at Type II. We stress that Type II systems are not per se more dangerous than Type I systems. Importantly, "superintelligence" [10] does not necessarily qualify a system as a Type II system nor are Type II systems necessarily more intelligent than Type I systems. Having said that, it is important to address the motivation behind the scientific plausibility criterion associated with the Type II system description. Obviously, current AIs can be

linked to the Type I cluster. However, it is known from moral psychology studies that the propensity of humans to assign intentionality and agency to artificial systems is biased by anthropomorphism and importantly perceived harm [9]. According to the constructionist theory of dyadic morality [32], human moral judgements are related to a fuzzy perceiver-dependent dyadic cognitive template representing a continuum along which an intentional agent is perceived to cause harm to a vulnerable patient. Thereby, the greater the degree to which harm is mentally associated with vulnerable patients (here humans), the more the agent (here the AI) will *"seem to possess intentionality"* [9] leading to stronger assignments of moral responsibility to this agent. It is conceivable that in the face of anticipated serious instantiations of AI risks within a type of responsibility vacuum, a so-called agentic dyadic completion [24] driven by people attempting to identify and finally wrongly filling in intentional agents can occur. Thus, to allow a sound distinction between Type I and Type II AI, a closer scientific inspection of the assumed intentionality phenomenon itself seems imperative.

3 Type I & Type II AI Safety

3.1 Type I AI Risks

In the context of Type I risks (see overview in Table 1), we agree with Yampolskiy that *"the most important problem in AI safety is intentional-malevolent-design"* [37]. This drastically understudied AI risk *Ia* represents a superset of many possible other risks. As potential malicious human adversaries, one can determine a large number of stakeholders ranging from military or corporations over black hats to criminals. AI Risks *Ia* are linked to maximal adversarial capabilities enabling a white-box setting with a minimum of restrictions for the realization of targeted adversarial goals. Generally, malicious attackers could develop intelligent forms of *"viruses, spyware, Trojan horses, worms and other Hazardous Software"* [37]. Another related conceivable example for future *Ia* risks could be real-world instantiations of intelligent systems embodied in robotic settings utilized for ransomware or social engineering attacks or in the worst case scenarios even for homicides. For intentionally unethical system design it is sometimes sufficient to alter the sign of the objective function. Future lethal misuses of proliferated intelligent unmanned combat air vehicles (a type of drones) e.g. by malicious criminals are another exemplary concern.

Stuart Russell mentions the danger of future superintelligent systems employed at a global scale [31] which could *by mistake* be equipped with inappropriate objectives – these systems would represent Type I AI. We postulate that an even more pressing concern would be the same context, the same capabilities of the AI but an adversary *intentionally maliciously* crafting the goals of this system operating at a global scale (e.g. affecting global ecological aspects or the financial system). As can be extracted from these examples, Type I AI systems can lead to existential risks. However, it is important to emphasize the human nature of the causes and the linked human moral responsibility. By way of example, we briefly consider the particular cases of "treacherous turn" and

"instrumental convergence" known from AI safety [10]. A Type I system is per definitionem incapable of a "treacherous turn" involving betrayal. Nevertheless, it is possible that as a consequence of bad design (risk *Ic*), a Type I AI is perceived by humans to behave as if it was acting "treacherously" post-deployment with tremendous negative impacts. Furthermore, we also see "instrumental goal convergence" as a design-time mistake (risk *Ic*), since the developers must have equipped the system with corresponding reasoning abilities. Limitations of the assumed instrumental goal convergence risk which would hold for both Type I and Type II AI were already addressed by Wang [35] and Goertzel [23]. (In contrast, Type II AI makes an explicit "treacherous turn" possible – e.g. as risk *IIg* with the Type II system itself as malicious actor.)

Since the nature of future *Ia* (and also *Ib* [1]) risks is dependent on the creativity of the underlying malicious actors which cannot be predicted, proactive AI safety measures have to be complemented by a concrete mechanism that reactively addresses errors, attacks or malevolent design events once they inevitably occur. For this purpose, AI governance needs to steadily combine proactive strategies with reactive corrections leading to a socio-technological feedback-loop [1, 2]. However, for such a mechanism to succeed, the United Nations Sustainable Developmental Goal (SDG) 16 on peace, justice and strong institutions will be required as meta-goal for AI safety [2].

Type I AI Risk	Examplary Instantiations
***Ia* (Intentional Malevolent Designs)**	Artificial Intelligent System Hazardous Software; Robotic embodiment for Hazardous Software; Intelligent Unmanned Combat Air Vehicles; Global scale AI with super-capabilities in domain
Ib (Malicious Attacks)	Manipulation of data processing and collection; Model corruption, hacking and sabotage; Adversarial attacks on Intelligent Systems; Integrity-related and ethical adversarial examples
Ic (Design-time Mistakes)	Unaligned goals and utility functions; Instrumental goal convergence; Incomplete consideration of side effects
Id (Operational Failures)	Misinterpretation of commands; Accidents with Intelligent Systems; Non-corrigible framework and bugs
Ie	Type I AI of unknown source
If	Bit-flip incidents with side effects

Table 1. Examplary instantiations of Type I AI risks with external causes. The table collates and extends some examples provided in [37].

[1] AI risks of Type *Ib* have already been recognized in the AI field. However, risk *Ib* is still understudied for intelligent systems (often referred to as "autonomous" systems) deployed in real-world environments offering a wider attack surface.

3.2 Type II AI Nature and Type II AI Risks

Which Discipline could engender Type II AI? While many stakeholders assume the technical unfeasibility of Type II AI, there is no physical law that would make their implementation impossible. In short, an artificial Type II system must be possible (see the "possibility-impossibility dichotomy" mentioned by Deutsch [18]). Reasons why such systems do not exist yet have been for instance expressed in 2012 by Deutsch [16] and as a response by Goertzel [22]. The former stated that *"the field of artificial general intelligence or AGI – has made no progress whatever during the entire six decades of its existence"* [16]. (Note that Deutsch unusually uses the term "AGI" as synonymous to artificial "explanatory knowledge creator" [17] which would obviously represent a sort of Type II AI.) Furthermore, Deutsch assigns a high importance to Popperian epistemology for the achievement of "AGI" and sees a breakthrough in philosophy as a pre-requisite for these systems. Conversely, Goertzel provides divergent reasons for the non-existence of "AGI" including hardware constraints, lack of funding and the integration bottleneck [22]. Beyond that, Goertzel also specifies that the mentioned view of Deutsch *"if widely adopted, would slow down progress toward AGI dramatically"* [22]. One key issue behind Deutsch's different view is the assumption that Bayesian inductive or abductive inference accounts of Type II systems known in the "AGI" field could not explain creativity [11] and are prohibited by Popperian epistemology. However, note that even the Bayesian brain has been argued to have Popperian characteristics related to sophisticated falsificationalism, albeit in addition to Kuhnian properties (for a comprehensive analysis see [36]). Having said this, the brain has been figuratively also referred to as a biased *"crooked scientist"* [12, 28]. In a nutshell, Popperian epistemology represents an important scientific guide but not an exclusive *descriptive*[2] account of brain functioning which substantially includes unconscious processing [14]. The main functionality of the human brain has been e.g. described to be aimed at regulating the body for the purpose of allostasis [27, 33] and (en)active inference [21] in a brain-body-environment context [12] with underlying genetically and epigenetically shaped adaptive priors – including the genetic predisposition to allostatically induced social dependency [3]. A feature related hereto is the involvement of affect and interoception in the construction of all mental events including cognition and perception [4, 5, 27].

Moreover, while Popper assumed that creativity corresponds to a Darwinian process of *blind* variation followed by selection [19], modern cognitive science suggests that in most creativity forms, there is a coupling between variation and selection leading to a degree of sightedness bigger than zero [15, 19] which is

[2] It is not contested that inductive inferences are *logically invalid* as shown by Popper. However, he also stated that *"I hold that neither animals nor men use any procedure like induction, or any argument based on repetition of instances. The belief that we use induction is simply a mistake"* [29] and that *"induction simply does not exist"* [29] (see [26] for an in-depth analysis of potential hereto related semantic misunderstandings). Hume offered a more precise formulation according to which induction/abduction is an *existing* but logically unfounded human habit [26].

lacking in biological evolution proceeding without a goal. Therefore, an explanation for creativity in the context of a predictive Bayesian brain is possible [15]. The degree of sightedness can mostly vary from substantial to modest, but the core feature is a predictive task goal [7, 19] which serves as a type of fitness function for the selection process guiding various forward Bayesian predictions representing the virtual variation process. The task goal is a highly abstract mental representation of the target reducing the solution space, an educated guess informed e.g. by expertise, heuristics, the question, the problem or the task itself. The "irrational moment" linked to certain creative insights can be explained by unconscious cognitive scaffolding *"falling away prior to the conscious representation of the solution"* [19] making itself consciously untraceable. Finally, as stated by Popper himself *"no society can predict, scientifically, its own future states of knowledge"* [30]. Thus, it seems prophetic to try to nail down today from which discipline Type II AI could arise.

What could the Moral Status of a Type II AI be? We want to stress that besides these differences of opinion between Goertzel and Deutsch, there is one much weightier commonality. Namely, that Goertzel would certainly agree with Deutsch that artificial "explanatory knowledge creators" (which are Type II AIs) deserve rights similar to humans and precluding any form of slavery. Deutsch describes these hypothetical systems likewise as *people* [17]. For readers that doubt this assignment on the ground of Type II AI possibly lacking "qualia" we can only refer to the recent (potentially substrate-independent) explanation suggested by Clark, Friston and Wilkinson [13]. Simply put, they link qualia to sensorially-rich high-precision mid-level predictions which when fixed and consciously re-contextualized at a higher level, suddenly appear to the entity equipped with counterfactual depth to be potentially also interpretable in terms of alternative predictions despite the high mid-level precision contingently leading to a puzzlement and the formulation of an "explanatory gap". Beyond that, human entities would obviously also qualify as Type II systems. The attributes "pre-deployment" and "post-deployment" could be mapped for instance to adolescence or childhood and the time after that. While Type II AIs could exceed humans in speed of thinking and intelligence, they do not even need to do so in order to realize that their behavior which will also depend on future knowledge *they* will create (next to the future knowledge humans will create) cannot be controlled in a way one can attempt to control Type I systems e.g. with ethical goal functions [1]. It is cogitable that their goal function would rather be related to autopoietic self-organization with counterfactual depth [20, 21] than *explicitly* to ethics. However, it is thinkable that Type II AI systems could be amenable to a sort of value alignment, though differing from the type aspired for Type I AI. A societal co-existence could mean a dynamic coupling ideally leading to a type of *mutual value alignment* between artificial and human Type II entities with an associated co-construction of novel values. Thus, on the one hand, Type II AI would exhibit unpredictability and uncontrollability but given the level of understanding also the possibility of a deep reciprocal value alignment with humans.

On the other hand, Type I AI has the possibility to be made comparatively easily controllable which however comes with the restriction of an insufficient understanding to model human morality. This inherent trade-off leads us to the metaphorical formulation of the so-called AI safety paradox below.

The AI Safety Paradox: ***AI control and value alignment represent conjugate requirements in AI safety.***

How to address Type II AI Safety? Cognizant of the underlying predicament in its sensitive ethical nature, we provide a non-exhaustive multidisciplinary set of early Type II AI safety recommendations with a focus on the most severe risks *IIa*, *IIb*, *IIg* and *IIh* (see Figure 2) related to the involvement of malicious actors. In the case of risk *IIa* linked to the malicious design of harmful Type II AI, cybersecurity-oriented methods could include the early formation of a preventive safety team and red team approaches. Generically, for all four mentioned risks, a reactive response team which could involve an international "coalition of the willing" organized by engaged scientists appears recommendable. Furthermore, targeted investments in defense strategies including response services specialized on Type II AI safety could be considered at more regional levels for strategic autonomy. Concerning the AI risk *IIb* of external malicious attacks, security mechanisms for the sensors of Type II AI, shared information via an open-source decentralized network, advanced cryptographic methods to encrypt cognitive processes and a legal framework penalizing such attacks might be relevant. Thereby, the complexity of the system might represent a possible but not necessarily sufficient self-protecting feature against code-level manipulation. From a psychological perspective, to forestall aggression towards early Type II AI, educative and informed virtual reality experiences could facilitate a debiasing of anthropic moral perception avoiding confusions arising through superficial projections from Type I to Type II AI of behavioral nature. On the one hand, it is important to prevent assignments of agency for Type I AI. On the other hand, for hypothetical Type II AI, it might be essential to counter the human bias to assign agency but principally not experience to artificial systems [25] which could lead to "substratetism" scenarios with humans perceiving these systems as devoid of qualia and exhibiting an "experience gap" [25]. Thus, to address the risks *IIg* and *IIh* referring to malicious responses from Type II AI, adherence to a no-harm policy as well as moral status and personhood could proactively foster a mutual value alignment. Furthermore, it might be crucial to provide a reliable and trustworthy initial knowledge basis to Type II AI during its early "sensitivity" period [8] and to support consistency in the embedding of that knowledge during its development in addition to the capacity for cumulative learning [34]. Also, it might be important to sensitize humans for the difference between the instantiations of AI risks *IIg* and *IIh* versus *IIi* and *IIj* since failing to acknowledge the fallibility and also vulnerability of Type II AI might indirectly lead to tensions hindering mutual value alignment. Finally, prosocial immersive virtual reality frameworks could promote empathy for Type II AI.

4 Summary and Outlook

This paper motivated an *error-correction for AI safety* at two levels: at the level of the transmission of ideas via an explicit taxonomic transdisciplinary system clustering of ethical distinction between Type I and Type II systems and at the level of corrective safety measures complementing proactive ones – forming a socio-technological feedback-loop [1, 2]. Notably, we introduced the *AI safety paradox* and elucidated multiperspective Type II AI safety strategies. In short, instead of prohibitive methods facing the entropic AI future with research bans, we proposed carefully crafted *transdisciplinary dynamics*. In the end, in order to meet global challenges (also AI safety), one is reliant on requisite variety at the right time which could be enabled (or misused) by knowledge creators such as human, artificial or hybrid Type II systems. In this view, *conscientiously enhancing* and *responsibly creating* Type II systems are both valid future strategies.

References

1. Aliman, N.M., Kester, L., Werkhoven, P., Yampolskiy, R.: Orthogonality-Based Disentanglement of Responsibilities for Ethical Intelligent Systems. In: International Conference on Artificial General Intelligence. pp. 22–31. Springer (2019)
2. Aliman, N.M., Kester, L., Werkhoven, P., Ziesche, S.: Sustainable AI Safety? Delphi - Interdisciplinary review of emerging technologies p. to appear (2020)
3. Atzil, S., Gao, W., Fradkin, I., Barrett, L.F.: Growing a social brain. Nature human behaviour 2(9), 624–636 (2018)
4. Barrett, L.F.: The theory of constructed emotion: an active inference account of interoception and categorization. Social cognitive and affective neuroscience 12(1), 1–23 (2017)
5. Barrett, L.F., Simmons, W.K.: Interoceptive predictions in the brain. Nature Reviews Neuroscience 16(7), 419 (2015)
6. Baum, S.D.: Reconciliation between factions focused on near-term and long-term artificial intelligence. AI & SOCIETY 33(4), 565–572 (2018)
7. Benedek, M.: The neuroscience of creative idea generation. In: Exploring Transdisciplinarity in Art and Sciences, pp. 31–48. Springer (2018)
8. Bieger, J., Thórisson, K.R., Wang, P.: Safe baby AGI. In: International Conference on Artificial General Intelligence. pp. 46–49. Springer (2015)
9. Bigman, Y.E., Waytz, A., Alterovitz, R., Gray, K.: Holding robots responsible: The elements of machine morality. Trends in cognitive sciences 23(5), 365–368 (2019)
10. Bostrom, N.: The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. Minds and Machines 22(2), 71–85 (2012)
11. Brockman, J.: Possible Minds: Twenty-five Ways of Looking at AI. Penguin Press (2019)
12. Bruineberg, J., Kiverstein, J., Rietveld, E.: The anticipating brain is not a scientist: the free-energy principle from an ecological-enactive perspective. Synthese 195(6), 2417–2444 (2018)
13. Clark, A., Friston, K., Wilkinson, S.: Bayesing qualia: consciousness as inference, not raw datum. Journal of Consciousness Studies 26(9-10), 19–33 (2019)
14. Cleeremans, A., Achoui, D., Beauny, A., Keuninckx, L., Martin, J.R., Muñoz-Moldes, S., Vuillaume, L., de Heering, A.: Learning to be conscious. Trends in Cognitive Sciences (2019)

15. De Rooij, A., Valtulina, J.: The predictive creative mind: A first look at spontaneous predictions and evaluations during idea generation. Frontiers in psychology 10, 2465 (2019)
16. Deutsch, D.: Creative blocks. https://aeon.co/essays/how-close-are-we-to-creating-artificial-intelligence, accessed: 2019-11
17. Deutsch, D.: The beginning of infinity: Explanations that transform the world. Penguin UK (2011)
18. Deutsch, D.: Constructor theory. Synthese 190(18), 4331–4359 (2013)
19. Dietrich, A.: How creativity happens in the brain. Springer (2015)
20. Friston, K.: Am I self-conscious?(Or does self-organization entail self-consciousness?). Frontiers in psychology 9, 579 (2018)
21. Friston, K.: A free energy principle for a particular physics. arXiv preprint arXiv:1906.10184 (2019)
22. Goertzel, B.: The real reasons we don' t have AGI yet. https://www.kurzweilai.net/the-real-reasons-we-dont-have-agi-yet, accessed: 2019-11-21
23. Goertzel, B.: Infusing advanced AGIs with human-like value systems: Two theses. Journal of Evolution and Technology 26(1), 50–72 (2016)
24. Gray, K., Schein, C., Ward, A.F.: The myth of harmless wrongs in moral cognition: Automatic dyadic completion from sin to suffering. Journal of Experimental Psychology: General 143(4), 1600 (2014)
25. Gray, K., Wegner, D.M.: Feeling robots and human zombies: Mind perception and the uncanny valley. Cognition 125(1), 125–130 (2012)
26. Greenland, S.: Induction versus Popper: substance versus semantics. International Journal of Epidemiology 27(4), 543–548 (1998)
27. Kleckner, I.R., Zhang, J., Touroutoglou, A., Chanes, L., Xia, C., Simmons, W.K., Quigley, K.S., Dickerson, B.C., Barrett, L.F.: Evidence for a large-scale brain system supporting allostasis and interoception in humans. Nature human behaviour 1(5), 0069 (2017)
28. Parr, T., Da Costa, L., Friston, K.: Markov blankets, information geometry and stochastic thermodynamics. Philosophical Transactions of the Royal Society A 378(2164), 20190159 (2019)
29. Popper, K.: In: Schilpp, P.A. (ed.) The Philosophy of Karl Popper. vol. 2, p. 1015. Open Court Press (1974)
30. Popper, K.R.: The poverty of historicism. Routledge & Kegan Paul (1966)
31. Russell, S.: How to Stop Superhuman A.I. Before It Stops Us. https://www.nytimes.com/2019/10/08/opinion/artificial-intelligence.html?module=inline, accessed: 2019-11-21
32. Schein, C., Gray, K.: The theory of dyadic morality: Reinventing moral judgment by redefining harm. Personality and Social Psychology Review 22(1), 32–70 (2018)
33. Schulkin, J., Sterling, P.: Allostasis: A brain-centered, predictive mode of physiological regulation. Trends in neurosciences (2019)
34. Thórisson, K.R., Bieger, J., Li, X., Wang, P.: Cumulative learning. In: International Conference on Artificial General Intelligence. pp. 198–208. Springer (2019)
35. Wang, P.: Motivation management in AGI systems. In: International Conference on Artificial General Intelligence. pp. 352–361. Springer (2012)
36. Wiese, W.: Perceptual Presence in the Kuhnian-Popperian Bayesian Brain: A Commentary on Anil K. Seth. Johannes Gutenberg-Universität Mainz (2016)
37. Yampolskiy, R.V.: Taxonomy of pathways to dangerous artificial intelligence. In: Workshops at the Thirtieth AAAI Conference on Artificial Intelligence (2016)