Hybrid Strategies Towards Safe "Self-Aware" Superintelligent Systems

Nadisha-Marie Aliman[1](✉) and Leon Kester[2]

[1] University of Stuttgart, Stuttgart, Germany
nadishamarie.aliman@gmail.com
[2] TNO Netherlands, The Hague, Netherlands

Abstract. Against the backdrop of increasing progresses in AI research paired with a rise of AI applications in decision-making processes, security-critical domains as well as in ethically relevant frames, a large-scale debate on possible safety measures encompassing corresponding long-term and short-term issues has emerged across different disciplines. One pertinent topic in this context which has been addressed by various AI Safety researchers is e.g. the AI alignment problem for which no final consensus has been achieved yet. In this paper, we present a multidisciplinary toolkit of AI Safety strategies combining considerations from AI and Systems Engineering as well as from Cognitive Science with a security mindset as often relevant in Cybersecurity. We elaborate on how AGI "Self-awareness" could complement different AI Safety measures in a framework extended by a jointly performed Human Enhancement procedure. Our analysis suggests that this hybrid framework could contribute to undertake the AI alignment problem from a new holistic perspective through security-building synergetic effects emerging thereof and could help to increase the odds of a possible safe future transition towards superintelligent systems.

Keywords: Self-awareness · AI Safety · Human enhancement
AI alignment · Superintelligence

1 Introduction

Being a topic of major importance in AI Safety research, AI alignment – which is often interchangeably used with the term of value alignment – has been analyzed from diverse point of views and incorporates a variety of research subareas many of which were reviewed by Taylor et al. [29]. Two highly relevant approaches in the realization of AI alignment the authors considered in this context are *value specification* and *error tolerance* which were both introduced by Soares and Fallenstein [28]. In order to do justice to these two distinct issues, Taylor et al. postulate that *"we can do research that makes it easier to specify our intended goals as objective functions"* concerning the first and *"we can do research aimed at designing AI systems that avoid large side effects and negative incentives,*

M. Iklé et al. (Eds.): AGI 2018, LNAI 10999, pp. 1–11, 2018.
https://doi.org/10.1007/978-3-319-97676-1_1

even in cases where the objective function is imperfectly aligned" concerning the latter. We take these high-level considerations alongside additional multidisciplinary observations as point of departure and apply a more abstract and holistic analysis than many prior papers have utilized in this particular context to identify solution approaches. For instance, we see the need for "self-awareness" in AI systems for reasons such as safety, effectiveness, transparency or explainability just as such a functionality is required from the perspective of Systems Engineering for the effectiveness and safety of advanced models. Beyond that, we agree that methods inspired from Cybersecurity practices [20] could provide a valuable support for AI Safety including the safety of self-aware AGIs. Furthermore, we also focus on the human factor in the AGI development and suggest to make allowance for human cognitive constraints in AI Safety frameworks taking a perspective jointly considering ethical aspects.

In the next Sect. 2, we posit that a (yet to be defined) "self-awareness" functionality might beside other benefits account for an enhanced error tolerance within a future human-level AGI model and might indirectly facilitate the value or goal specification process. Thereafter, in Sect. 3, we suggest that a self-aware AGI that should be deployed in a real-world environment will have to be supplemented by additional AI Safety measures including for instance an AGI Red Teaming approach in order to maintain a high error tolerance level. In Sect. 4, we analyse how AGI developers could proficiently face the problem of adequate value specification in the first place, which could interestingly imply the need for an enhancement of human "self-awareness" to a certain extent with respect to the goal to identify the values humans really intend on the one hand and regarding the aim to subsequently encode this values into prioritized goals a self-aware AGI will have to adhere to on the other hand. Finally, in the last Sect. 5, we reflect upon this set of hybrid strategies as an interwoven entirety, consider its possible ethical implications and place it in the context of a hypothetically thereof emerging type of superintelligence.

2 Self-Awareness

While the notion of "self-awareness" which is often used in the context of concepts like "self-conciousness", "self-control" or "self-reference" is not in the focus of classical AI research, it is considered to be one of the key elements out of the crucial competency areas for Human-Level General Intelligence according to many AGI researchers (as investigated by Adams et al. [1]) and the notion itself or related terms have been considered in some ways within various AGI designs (e.g. in [5,6,12,13,18,25,31,32]). However, the relevancy of AGI self-awareness from the perspective of AI Safety remains a poorly studied topic, even though the omission of such a functionality in an AGI architecture might lead to far-reaching implications in the future in regard to the safety of this system if deployed in a dynamic real-world environment. Given that a definition of this relatively abstract term is controversial and nontrivial, we will in the following first provide a simple technically oriented definition of AGI self-awareness – for which we do not claim any higher suitability in general, but which is specifically conceptualized for our line of argument – and then subsequently elucidate the reasons for its crucial importance in AI Safety frameworks.

The definition is inspired by Systems Engineering practices with applications to diverse types of dynamic systems as e.g. adapted by Kester et al. [14,15] or van Foeken et al. [10] and is not restricted to the choice of any particular AGI architecture provided that the AGI acts in a not further defined goal-oriented manner, possesses sensors and actuators as well as the ability to somehow communicate with human entities. For clarity, when we refer to an AGI exhibiting *self-awareness* in this work, we explicitly mean an AGI which is able to independently perform *self-assessment* and *self-management*, whereby self-assessment designates a set of processes enabling the AGI to determine the performance of its various functions with respect to its goals (e.g. for associated physical instances, internal cognitive processes, own abilities, own resources,...) by itself and self-management the capability to adapt its behavior in the real-world on its own in order to reach its goals based on the information collected through self-assessment. In addition, the AGI is presupposed to be able to communicate the insights obtained after having performed self-assessment and the choices made in the self-management step to specified human entities.

In the following, we collate some possible highly relevant advantages for a self-awareness functionality within an AGI architecture from the perspective of AI Safety:

- *Transparency:* Through the ability of a self-aware AGI to allow important insights into its internal processes to its designers, it by design does not correspond to a "black-box" system as it is the case for many contemporary AI architectures. The resulting transparency presents a valuable basis for effective AI Safety measures.
- *Explainability:* Since the AGI performs self-management on the basis of a transparent self-assessment, its decision-making process can be independently documented and communicated, which might increase the possibility for humans to extract helpful explanations for the actions of the AGI.
- *Trustworthiness:* An improved AGI explainability might increase its trustworthiness and acceptance from a human perspective, which might in turn offer more chances to test the self-aware AGI in a greater variety of real-world environments and contexts.
- *Controllability:* Through the assumed communication ability of the AGI, a steady feedback loop between human entities and the AGI might lead to an improved human control offering many opportunities for testing and the possibility to proactively integrate more AI Safety measures. More details on possible proactive measures are provided in the next Sect. 3.
- *Fast Adaptation:* Self-awareness allows for faster reactions and adaptations to changes in dynamic environments even in cases where human intervention might not be possible for temporal reasons which allows for an improved error tolerance and security. Unwanted scenarios might be more effectively avoided in the presence of negative feedback from the environment.

- *Cost-Effectiveness:* There is often a tradeoff between security and cost-effectiveness, however a self-aware system is inherently more cost-effective for instance due to the better traceability of its errors, the facilitated maintainability through the transparency of its decision-making processes or because the system can adapt itself to optimal working in any situation, while lacking any obvious mechanism which might in exchange lower its security level – by what a double advantage arises.
- *Extensibility*: Finally, a self-aware AGI could be extended to additionally for instance contain a model of human cognition which could consider human deficiencies such as cognitive constraints, biases and so on. As a consequence, the AGI could adapt the way it presents information to human entities and consider their specific constraints to maintain a certain level of explainability.

However, after having compiled possible advantages AGI self-awareness could offer to AI Safety, it is important to note that up to now, it was not specified on what basis the goals of the self-aware goal-oriented AGI are crafted in the first place. Moreover, the odds that a self-aware AGI spawns many of the mentioned desirable properties are even largely dependent on the quality of the goals assigned to it and it is thus clear that self-awareness taken alone is far from representing a panacea for AI Safety, since it does not per se solve the underlying goal alignment problem. Nonetheless, we argue that AGI self-awareness represents a highly valuable basis for future-oriented AI Safety measures due to the vitally important advantages it could bring forth if combined with appropriate goals. In addition, AGI self-awareness might be able to itself facilitate the process of goal alignment through the interactive transparent framework suitable for tests in real-world environments it offers, whereby the selection of adequate goals clearly remains a highly debatable topic on its own. From our perspective, the therefore required goal function intrinsically reflecting desirable human values for a self-aware AGI could be stipulated by humans which would be specifically trained in interaction with that AGI and possibly ethically as well as cognitively enhanced on the basis of technological advances/scientific insights, since humanity at its current stage, seems to exhibit rather insufficient solutions for a thoughtful and safe future in conjunction with AGIs – especially when it comes to the possible necessity for an unambiguous formulation of human goals. We will further address the motivations for human enhancement to provide assistance during this mentioned process of goal selection in Sect. 4.

3 Proactive AI Safety Measures

After having depicted possible benefits as well as still unanswered implications in the context of a self-aware AGI, we now focus on crucial AI Safety measures which might be necessary in addition to avoid unintended harmful outcomes during the development phase and prevent risky scenarios after a subsequent deployment of such an AGI architecture. While the suggested methods would undoubtedly not guarantee an absolutely risk-free AGI, their indispensability to at least obtain a well tested architecture built with a certain security awareness which particularly also takes the possibility of intentionally malevolent actors [20] into account, seems however to prohibit their omission. Beyond that, it seems imperative to incorporate a type of simulations of undesirable scenarios while developing an AGI as a proactive rather than reactive approach, since the latter might be reckless given the extent of possible future consequences which could include a number of existential risks [7,20,30].

In the long run, further research on the following (unquestionably non-exhaustive and extendable) measures building on previous work and extending certain concepts could offer forward-looking hints in this regard:

- *Development Under Adversarial Assumptions:* Already during the AGI development phase, the developers should take into account the most important known types of e.g. integrity vulnerabilities that have been reported regarding other AIs in the past (this could include rather similar architectures, but importantly also cognitively less sophisticated AIs since it could represent a type of minimum requirement) and should not per default conjecture a benign environment. In a simplified scheme, assuming the development of an AGI starting nowadays, it should for instance among others be ascertained that none of the known adversarial methods to fool narrow AIs such as Deep Neural Networks [19] would also lead to a defective information processing of security-relevant kind if correspondingly corrupted inputs are presented to the sensors of the AGI at hand. Besides that, new types of A(G)I attacks and corresponding defense mechanisms should be actively ethically investigated. In this context a new subfield of study on "adversarial examples for AGIs" appears recommendable. While adversarial examples for narrow AIs are for instance associated with definitions such as *"inputs to machine learning models that an attacker has intentionally designed to cause the model to make a mistake"*[1], a corresponding analogy could be derived for AGIs. Ideally, the self-aware AGI itself could be trained in identifying situations susceptible to involve particular known safety threats.
- *AGI Red Team:* As it is the case in the context of security systems, developers tend to be biased towards emphasizing the robustness of their system and might additionally exhibit "blind spots" to existing vulnerabilities while implementing defense strategies [16], which is why realistic red team events offer an invaluable security tool in many Cybersecurity frameworks [22–24]. Red Teaming has recently as well be proposed by Brundage et al. [8] in the context of recommendations for an AI Safety framework covering short-term issues for the next 5 years. Similarly, an external AGI red team could in the long-term periodically perform real-world attack simulations after the deployment of an AGI, with the goal to identify certain types of possibly overlooked vulnerabilities to sophisticated attacks. The red team could for instance explicitly try to trigger unethical actions on the part of the AGI by placing it in unknown or unusual contexts. In these settings, the blue team would correspond to the AGI developers which are responsible for the defense design within the AGI architecture. Possibly, social engineering performed by

[1] Mentioned in: https://blog.openai.com/adversarial-example-research/.

the red team on the blue team could disclose biases underlying the AGI training or its architecture and facilitate the crafting of specific targeted attacks. It is to be expected that such red team exercises will contribute to strengthen the robustness and possibly even enhance the cognitive abilities of the AGI by providing the AGI developers with comprehensive hints on how to enhance the defense designs which could for instance be of meta-cognitive nature. The ultimate objective would be to achieve a state from which on the self-aware AGI has learned to automatically and independently run self-tests simulating such systematical adversarial attacks.

- *Regular Measurement of Cognitive Ability and Inhibition of Self-interest:* To maintain transparency and allow for a certain minimal monitoring of the AGI, it might be essential to be regularly aware of the level of cognitive ability it exhibits in order to customize the security measures. Besides classically proposed Turing Tests, one further interesting type of test is the recently proposed "test for detecting qualia" introduced by Yampolskiy [33] and based on visual illusions. Even if – from a philosophical point of view – it could be debatable whether the described test measures the presence of qualia itself, we suppose that it could provide invaluable cues to detect higher cognitive abilities as exhibited by an AGI, since just like human misperceptions (including e.g. optical illusions) can for instance help to better understand the mechanisms underlying the perception of humans in Cognitive Science, so could the analysis of AGI misperceptions analogously help to understand the internals of an AGI system. An automatic program could periodically test the AGI and generate an alarm in the case of "cognitive anomalies" indicating an unusual increase of cognitive capacity. This regular test could also be implemented as a self-test mechanism within the self-aware AGI architecture itself. However, an explicit protective mechanism that prevents the AGI from evolving any kind of harmful intrinsic goals out of self-interest should be additionally designed in order to obviate any undesirable takeoff scenario. A related core idea to prevent an AGI from evolving a type of misaligned self-interest has been described by Goertzel [11] in the context of his suggestion for a specifically designed "AI Nanny" developed with a pre-defined set of goals and encompasses for instance *"a strong inhibition against modifying its [the AI Nanny's] preprogrammed goals"* or *"a strong inhibition against rapidly modifying its general intelligence"*.

Yet, these strategies in combination with AGI self-awareness taken alone might not be sufficient given the human component in the development of the AGI entailing a wide array of undesirable ethical, cognitive and evolutionary biases.

4 Human Enhancement

Whereas in the context of the value alignment problem, the focus is often set on how future AGIs could optimally learn values from human agents be it for instance by imitation or by predefined ethical goals, a jointly performed technology-supported learning approach for human agents to enhance their cognitive abilities and ethical frameworks in order to be able to develop improved capabilities qualifying them to more competently deal with this highly relevant problem in the first place, remains an under-explored topic. Given the large array of human deficiencies including for instance cognitive biases [34], unintentional unethical behavior [26] or limitations of human information processing which could be considered as major handicaps in succeeding to solve the AI alignment problem, the approach to extend the abilities of humans in charge of developing an ethical AGI by science and technology emerges as auspicious strategy, however certainly not without reservations.

We postulate that the following two complementary types of human enhancement could be decisive to ameliorate the value specification abilities of humans improving the odds to succeed in AI alignment:

- *Ethical Enhancement:* One prominent subproblem of goal alignment can be simply described as to make the AI learn human goals [30]. For this purpose, humans obviously need to be first aware of the values they really intend to implement in order to encode them as a factual set of prioritized goals within an AGI model. Similarly, as stated in [3], humans need to become better "ethical regulators" (e.g. of themselves and of AIs) in an era which will be more and more shaped by AI. This task might inter alia require a better type of "self-assessment" on the part of humans – especially with regard to their own concrete ethical preferences, abilities and constraints. To improve the required human ethical self-assessment for the development of safe AGIs, developers should consider a dynamic multifarious science-based ethical framework which could for instance encompass debiasing training [17] as well as methods from behavioral ethics [9] and could in the future even include a type of AGI-assisted debiasing training where the same self-aware AGI which is periodically checked for safety could e.g. act as "teacher" in game settings providing a personalized feedback to its developers which could be expanded to a testing of acquired ethically relevant skills. Additionally, the group formation of the AGI developers itself should ideally reflect a synergetic heterogeneity of worldviews to fend off inequality and unnecessary biases at the core of the goal selection process.
- *Cognitive Enhancement:* Some decades ago, the cybernetics pioneer Ross Ashby expressed the following train of thought [4]: *"[...] it is not impossible that what is commonly referred to as "intellectual power" may be equivalent to "power of appropriate selection". [...] If this is so, and as we know that power of selection can be amplified, it seems to follow that intellectual power, like physical power, can be amplified."* Even if this statement might still reflect a controversial issue and human enhancement technologies are still in their infancy, expected progresses in areas such as Nanorobotics, Bionics, Biotechnology, Brain-Computer Interface research or the newly arisen field of Cyborg Intelligence integrating *"the best of both machine and biological intelligences"* [27] might lead to considerably extended possibilities for cognitive enhancement in the foreseeable future. Transferring the term used

> in Ashby's statement to a different context, we argue that (possibly AGI-assisted) methods to increase the human "power of appropriate *goal* selection" within the framework of AGI development given the ethical values agreed upon while supported by preceding ethical enhancement procedures, represent an essential future research direction to be pursued for AI Safety reasons. For this purpose, one could first experimentally start with presently rather primitive and clearly not sufficient enhancement concepts such as mental training, HMI tools, neurofeedback, non-invasive brain stimulation methods, multi-mind BCIs for decision-making or nootropics. Later on, a reasonable priority for a self-aware AGI might even be to generate methods facilitating human cognitive enhancement and develop concepts where if procurable the AGI augments rather than surrogates human entities initiating a bidirectional learning framework. Besides that, the group composition of AGI developers should ideally promote multidisciplinarity in order to reduce the occurrences of AI Safety relevant blind spots in the development phase and should comprise numerous partcipants with diverse research backgrounds.

While it should be clear that human enhancement pathways (such as through brain-machine collaboration) cannot guarantee the prevention of an occurring unethical AGI [2], not to perform human enhancement does not guarantee it either. Furthermore, the abstention from ethical human enhancement also does not necessarily prevent the performance of unethical human enhancement by malevolent actors at a later stage. Therefore, we argue that the early practice of human enhancement for ethical purposes like the improvement of the value specification process for AI alignment, might increase the odds of a resulting ethical AGI and could even in the long-term facilitate the detection of potential unethical AGI development or unethical human enhancement through the bundled cognitive and ethical abilities that could emerge out of the suggested bidirectional framework of mutual enhancement.

5 Conclusion and Future Prospects

In this work, we postulated that AGI self-awareness represents a highly valuable functionality from the perspective of AI Safety as it might be helpful for the error tolerance subtask of AI alignment as well as indirectly for value specification and provides many advantages such as transparency or explainability. We then introduced a number of proactive AI Safety measures including AGI Red Teaming which could be necessary in addition to the self-awareness functionality to maintain security and which might be beneficial for the error tolerance subproblem. We set forth that the described framework alone might not be sufficient due to the ethical and cognitive constraints AGI developers exhibit as human beings and proposed a jointly performed inter alia AI-assisted ethical as well as cognitive enhancement procedure to support the goal selection process. We do not claim that the described hybrid framework represents a complete approach warranting the safety of the AGI or of a therefrom emerging superintelligence, but argue that it might underpin the importance of a multidisciplinary approach to AI Safety and motivate a new useful holistic perspective on the complex problem of AI alignment which might in turn shape future developments towards a beneficial form of superintelligence (be it of human, artificial or hybrid nature). Finally, we stress that possible future research on self-aware AGIs as well as research on ethical and cognitive enhancement for AI Safety should not be reserved to stakeholders like corporations, the military or a presumed elite group of AGI developers, but be instead performed open-source and shared across diverse communities for the benefit of mankind. Moreover, a science-based debate on the implications of a conjectured technological singularity (which is not bounded to necessarily emerge from an AGI [21]) should be encouraged and existential risks through superintelligence should be thoroughly taken into consideration – especially regarding scenarios implying the presence of malicious actors [2,20].

References

1. Adams, S.S., et al.: Mapping the landscape of human-level artificial general intelligence. AI Magaz. **33**, 25–41 (2012)
2. Aliman, N.-M.: Malevolent cyborgization. In: Everitt, T., Goertzel, B., Potapov, A. (eds.) AGI 2017. LNCS (LNAI), vol. 10414, pp. 188–197. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-63703-7_18
3. Ashby, M.: Ethical regulators and super-ethical systems. In: Proceedings of the 61st Annual Meeting of the ISSS-2017 Vienna, Austria, vol. 2017 (2017)
4. Ashby, W.R.: An Introduction to Cybernetics. Chapman & Hall Ltd., New York (1961)
5. Baars, B.J., Franklin, S.: Consciousness is computational: the LIDA model of global workspace theory. Int. J. Mach. Conscious. **1**(01), 23–32 (2009)
6. Bach, J.: Principles of Synthetic Intelligence PSI: An Architecture of Motivated Cognition, vol. 4. Oxford University Press, Oxford (2009)
7. Bostrom, N.: Superintelligence: Paths, Dangers, Strategies (2014)
8. Brundage, M., et al.: The malicious use of artificial intelligence: forecasting, prevention, and mitigation. arXiv preprint arXiv:1802.07228 (2018)
9. Drumwright, M., Prentice, R., Biasucci, C.: Behavioral ethics and teaching ethical decision making. Decis. Sci. J. Innovative Educ. **13**(3), 431–458 (2015)
10. van Foeken, E., Kester, L., Iersel, M.: Real-time common awareness in communication constrained sensor systems. In: Proceedings of 12th International Conference on Information Fusion, FUSION 2009, Seattle, Washington, USA, pp. 118–125, 6–9 July 2009
11. Goertzel, B.: Should humanity build a global AI nanny to delay the singularity until its better understood? J. Conscious. Stud. **19**(1–2), 96–111 (2012)
12. Goertzel, B.: Characterizing human-like consciousness: an integrative approach. Procedia Comput. Sci. **41**, 152–157 (2014)
13. Goertzel, B.: A formal model of cognitive synergy. In: Everitt, T., Goertzel, B., Potapov, A. (eds.) AGI 2017. LNCS (LNAI), vol. 10414, pp. 13–22. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-63703-7_2
14. Kester, L., Ditzel, M.: Maximising effectiveness of distributed mobile observation systems in dynamic situations. In: 2014 17th International Conference on Information Fusion (FUSION), pp. 1–8. IEEE (2014)

15. Kester, L.J.H.M., van Willigen, W.H., Jongh, J.D.: Critical headway estimation under uncertainty and non-ideal communication conditions. In: Proceedings of 17th International IEEE Conference on Intelligent Transportation Systems (ITSC), pp. 320–327 (2014)
16. Mirkovic, J., et al.: Testing a collaborative DDoS defense in a red team/blue team exercise. IEEE Trans. Comput. **57**(8), 1098–1112 (2008)
17. Morewedge, C.K., Yoon, H., Scopelliti, I., Symborski, C.W., Korris, J.H., Kassam, K.S.: Debiasing decisions: Improved decision making with a single training intervention. Policy Insights Behav. Brain Sci. **2**(1), 129–140 (2015)
18. Nivel, E., et al.: Bounded recursive self-improvement. arXiv preprint arXiv:1312.6764 (2013)
19. Papernot, N., McDaniel, P., Sinha, A., Wellman, M.: Towards the science of security and privacy in machine learning. arXiv preprint arXiv:1611.03814 (2016)
20. Pistono, F., Yampolskiy, R.V.: Unethical research: how to create a malevolent artificial intelligence. In: Proceedings of 25th International Joint Conference on Artificial Intelligence (IJCAI-16). Ethics for Artificial Intelligence Workshop (AI-Ethics-2016) (2016)
21. Potapov, A.: Technological singularity: what do we really know? Information **9**(4), 99 (2018)
22. Rajendran, J., Jyothi, V., Karri, R.: Blue team red team approach to hardware trust assessment. In: 2011 IEEE 29th International Conference on Computer Design (ICCD), pp. 285–288. IEEE (2011)
23. Rege, A.: Incorporating the human element in anticipatory and dynamic cyber defense. In: IEEE International Conference on Cybercrime and Computer Forensic (ICCCF), pp. 1–7. IEEE (2016)
24. Rege, A., Obradovic, Z., Asadi, N., Singer, B., Masceri, N.: A temporal assessment of cyber intrusion chains using multidisciplinary frameworks and methodologies. In: 2017 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (Cyber SA), pp. 1–7. IEEE (2017)
25. Schmidhuber, J.: Gödel machines: fully self-referential optimal universal self-improvers. In: Goertzel, B., Pennachin, C. (eds.) Artificial General Intelligence, pp. 199–226. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-68677-4_7
26. Sezer, O., Gino, F., Bazerman, M.H.: Ethical blind spots: explaining unintentional unethical behavior. Curr. Opin. Psychol. **6**, 77–81 (2015)
27. Shi, Z., Ma, G., Wang, S., Li, J.: Brain-machine collaboration for cyborg intelligence. In: Shi, Z., Vadera, S., Li, G. (eds.) IIP 2016. IAICT, vol. 486, pp. 256–266. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-48390-0_26
28. Soares, N., Fallenstein, B.: Agent foundations for aligning machine intelligence with human interests: a technical research agenda. In: Callaghan, V., Miller, J., Yampolskiy, R., Armstrong, S. (eds.) The Technological Singularity. TFC, pp. 103–125. Springer, Heidelberg (2017). https://doi.org/10.1007/978-3-662-54033-6_5
29. Taylor, J., Yudkowsky, E., LaVictoire, P., Critch, A.: Alignment for advanced machine learning systems. In: Machine Intelligence Research Institute (2016)
30. Tegmark, M.: Life 3.0: Being Human in the Age of Artificial Intelligence. Knopf, New York (2017)
31. Thórisson, K.R.: A new constructivist AI: from manual methods to self-constructive systems. In: Wang, P., Goertzel, B. (eds.) Theoretical Foundations of Artificial General Intelligence, pp. 145–171. Springer, Paris (2012). https://doi.org/10.2991/978-94-91216-62-6_9

32. Wang, P., Li, X., Hammer, P.: Self in NARS, an AGI system. Front. Robot. AI **5**, 20 (2018)
33. Yampolskiy, R.V.: Detecting qualia in natural and artificial agents. arXiv preprint arXiv:1712.04020 (2017)
34. Yudkowsky, E.: Cognitive biases potentially affecting judgment of global risks. Glob. Catastrophic Risks **1**(86), 13 (2008)