Orthogonality-Based Disentanglement Of Responsibilities For Ethical Intelligent Systems

Nadisha-Marie Aliman[1], Leon Kester[2], Peter Werkhoven[1,2], and Roman Yampolskiy[3]

[1] Utrecht University, Utrecht, Netherlands
[2] TNO Netherlands, The Hague, Netherlands
[3] University of Louisville, Louisville, USA

Abstract. In recent years, the implementation of meaningfully controllable advanced intelligent systems whose goals are aligned with ethical values as specified by human entities emerged as key subject of investigation of international relevance across diverse AI-related research areas. In this paper, we present a novel transdisciplinary and Systems Engineering oriented approach denoted *"orthogonality-based disentanglement"* which jointly tackles both the thereby underlying control problem and value alignment problem while unraveling the corresponding responsibilities of different stakeholders based on the distinction of two orthogonal axes assigned to the problem-solving ability of these intelligent systems on the one hand and to the ethical abilities they exhibit based on quantitatively encoded human values on the other hand. Moreover, we introduce the notion of explicitly formulated *ethical goal functions* ideally encoding what humans *should* want and exemplify a possible class of "self-aware" intelligent systems with the capability to reliably adhere to these human-defined goal functions. Beyond that, we discuss an attainable transformative socio-technological feedback-loop that could result out of the introduced orthogonality-based disentanglement approach and briefly elaborate on how the framework additionally provides valuable hints with regard to the coordination subtask in AI Safety. Finally, we point out remaining crucial challenges as incentive for future work.

Keywords: Ethical Goal Function, Self-Awareness, AI Alignment, Control Problem, AI Coordination

1 Motivation

In the current both safety-critical and ethically relevant international debate on how to achieve a meaningful control of advanced intelligent systems that comply with human values [19], diverse solution approaches have been proposed that fundamentally differ in the way they would affect the future development of A(G)I research. In a nutshell, one could identify a set of four main clusters of conceptually different solution approaches for which one could advocate for by distinguishing between 1) *prohibitive*, 2) *self-regulative*, 3) *deontological* and 4) *utility-based* methods. While the prohibitive approach aims at restricting or

even banning the development of highly sophisticated AI until problems related to control and value alignment are solved in the first place, it seems highly unlikely to be put into practice especially in its most extreme forms and it is therefore not further considered in this paper. By contrast, option 2) implies the assumption that certain mechanisms (for instance specific market mechanisms or mechanisms inherent to certain types of A(G)I architectures) could allow for a more or less automatically emerging stability or desirability of the behavior as exhibited by intelligent systems. Furthermore, solution 3) classically considers the direct hard-coding of ethical values into AI systems for instance by encoding deontological values at design time [18], while in the case of the utility-based approach 4), one mostly foresees a human-defined utility function [24] quantitatively encoding human values.

This debate – especially on whether to prefer the solution approach 3) or 4) – is often strongly imprinted by particularly difficult to solve philosophical issues and the AI-related responsibilities of different involved stakeholders such as users, programmers, manufacturers and legislators appears to be only vaguely and therefore insufficiently definable. Against this backdrop, the need for a practicable technically oriented and at the same time forward-looking solution appears to be of urgent importance for a responsible future planning of a hybrid society in close conjunction with advanced AI systems.

2 Disentanglement Of Responsibilities

For reasons of safety, security, controllability, accountability and reliability, it can be assumed that it is in the interest of a democratic society to achieve a transparent division of responsibilities for the deployment of intelligent systems in diverse application areas. Thereby, the systems should act in accordance with ethical and legal specifications as formulated by the legislative power and allow for traceability in order to facilitate an assignment of responsibility by the judicial power. Consequently, we argue that the self-regulative solution 2) can be ruled out since it would lead to a heterogeneous set of different ethical frameworks implemented within different types of intelligent systems yielding highly complex entanglements especially with regard to responsibility assignments (e.g. among manufacturers, programmers, users and operators). Furthermore, as the problem solving ability of the intelligent systems increases, the severity of possible unintended effects, malicious attacks [6] or the development of intentionally crafted unethical systems [17] which could even induce existential risks seems to prohibit a laissez-faire approach. Thus, the remaining options are the deontological approach 3) and the utility-based solution 4) since both could be in theory implemented within a framework separating the responsibilities as described.

According to the orthogonality thesis by Bostrom [5], *"intelligence and final goals are orthogonal axes along which possible agents can freely vary"*. Though, the thesis is not uncontroversial for reasons comprising the fact that it does not address probabilities as postulated by Goertzel [10]. However, for the purpose of our specific argument, it is not necessary to consider the soundness of the

thesis, since we only presuppose that "there exists a type of AI architecture for which final goals and intelligence are orthogonal" which is self-evident considering utility maximizers [4] as classical examples epistomizing solution 4). From this, it trivially follows that formulating a goal function for a utility maximizer and designing the architecture of this agent are separable tasks. Building on that, we argue that the already existing practice of the legislative power having a say on the *what* goals to achieve as long as societal impacts are concerned and the manufacturers implementing the *how* in various contexts can be adapted to goal-oriented utility maximizers (albeit with certain reservations particularly on the nature of the architecture used) and can thus be pursued as postulated by Werkhoven et al. [23].

Apart from that, it is undoubtedly possible to think of a similar disentanglement of responsibilities in accordance with a solution of the type 3). However, for mostly technical reasons we will now illustrate, we do not consider a deontological framework in which lawful and ethical behavior is encoded for instance in ontologies [12] or directly in natural language as possible instantiation of our orthogonality-based disentanglement approach. First, the attempt to formulate deontological rules for every possible situation in a complex unpredictable real-world environment ultimately leads to a state-action space explosion [23] (it is thereby obvious that law does not represent a complete framework). To be able to handle the complexity of such environments and the complexity of internal states, the intelligent system needs to be run-time adaptive which cannot be achieved by using static rules. Second, since law is formulated in natural language which is inherently highly ambiguous at multiple linguistic levels, the intelligent system would have to either make sense of the legal material using error-prone Natural Language Processing techniques or in the case of the ontology-based approach, the programmers/manufacturers would have to first interpret law before encoding it which induces uncertainty and violates the desired disentanglement of responsibilities. Third, law leaves many legal interpretations open and entails tradeoffs and dilemmas that an intelligent system might encounter and would need to address leading to an unspecified assignment of responsibilities. Fourth, an update of laws will require a costly and laborious update of designs for every manufacturer. Fifth, a deontological approach with fixed rules cannot easily directly endorse a process in which progresses in AI could be efficiently used to transform society in a highly beneficial way enabling humans to overcome their cognitive and evolutionary biases and creating new possibilities to improve the foundations of society.

Having expounded why the deontological solution approach 3) is inappropriate for the central problem of disentangling responsibilities for the deployment of intelligent systems, we now elucidate how a properly designed solution 4) is able to avoid all mentioned disadvantages associated with solution 3). First, it might be possible to realize run-time adaptivity within utility maximizers by equipping them with a "self-awareness" functionality [1] (self-assessment, self-management and the ability to deliver explanations for actions to human entities) which we outline in Section 4. Moreover, deontological elements could be used as con-

straints on the utility function of such utility maximizers in order to selectively restrict the action or the state space. Second, by quantifying law within a publicly available ethical goal function as addressed in the next Section 3, one achieves an increased level of transparency. Third, through a utility function approach tradeoffs and dilemmas are more easily and comprehensibly solved. Thereby, for safety reasons, the utility functions can and should include context-sensitive and perceiver-dependent elements as integrated e.g. in augmented utilitarianism [2]. Fourth, updates of law are solely reflected in the ethical goal functions which leads to a more flexible and controllable task. Fifth, the use of such an ethical goal function approach opens up the opportunity for a society to actively perform an enhancement of ethical abilities which we depict in Section 5.

3 Ethical Goal Function And "What One *Should* Want"

A first step of crafting ethical goal functions could be for instance to start with the mapping of each relevant application domain of law to a specific utility function which quantifies the expected utility of the possible transitions of the world. For this purpose, the legislative has for instance to define the relevant components of each goal function and assign weights to each component, decide which parameters to consider for each component and identify possible underlying correlations. (It is thinkable that specific stakeholders might then while applying the goal function to their particular area of application, craft a lower-level customized mission goal function [8] for their specific mission goals which would however have to be compliant with the ethical goal function provided by the legislative.) The implementation of this strategy will require a relatively broad multidisciplinary knowledge by policy-makers or might require the practical collaboration with trained multidisciplinary researchers with expertise in e.g. AI and Systems Engineering.

One important feature of the proposed framework is the requirement of transparent human-readable goal functions that can be inspected by anyone which substantially facilitates accountability. In order to obtain a specification of a human-oriented goal function, different methods have been proposed including inverse reinforcement learning (IRL) [9] and reward modeling [16]. However, the IRL method comes with the main drawback of yielding ambiguous reward functions that could explain the observed behavior and within reward modeling, a black-box model is trained by a user in order to act as reward function for a reinforcement learning agent which violates both the transparency requirement of our approach and the disentanglement of responsibilities since it is the user that trains the reward model (and not a representation of society).

However, it is important to note, that as implicit so far, the goal functions would be rather specified based on *what humans want* and not necessarily on what humans *should* want from a scientific perspective, since it is known that humans exhibit biases for instance inherent to their neural systems [15], due to their evolutionary past of survival in small groups [23] or through ethical blindspots [20] which represent serious constraints to their ethical views. On

these grounds, the framework described in this paper is intended to be of transformative and dynamical nature and might enable the legislative to receive a quantitatively defined feedback from the environment, which in turn might foster the human-made evidence-based adjustment of the explicitly formulated ethical goal functions towards more scientifically sound assumptions.

Beyond that, as postulated by Harris [11], a *science* of morality which might enable humans to identify the peaks on the "moral landscape" which he described as *"a [hypothetical] space of real and potential outcomes whose peaks correspond to the heights of potential well-being and whose valleys represent the deepest possible suffering"* could represent a feasible general approach to solve moral issues. In the light of the aforesaid, one could attempt to in the long-term pursue research that facilitates the design of a scientifically grounded universal ethical goal function whose local optima will ideally be conceptually equivalent to the peaks of this hypothetical moral landscape potentially reflecting what humans *should* want. Another interesting point of departure to be mentioned in this context, has been introduced by Ziesche [26] who describes how the UN sustainable development goals already representing an international consensus and containing values such as well-being could be quantified to start to practically tackle the value alignment problem.

Note that Yudkowsky' s early idea of a coherent extrapolated volition [25] in the context of friendly AI which envisaged an AI maximizing the utility based on an extrapolation of what we *would* want *"if we knew more, thought faster, were more the people we wished we were, had grown up farther together"* while being relatively close to it, is though subtly different from our described concept of what we *should* want based on a scientifically grounded ethical goal function, since an improvement of our problem solving ability does not necessarily improve our ethical abilities nor does *"the people we wished we were"* necessarily corresponds to a more ethical version of ourselves on average. Moreover, there is no reason to assume that human values would necessarily converge to ethical values if they *"had grown up farther together"*. However, as will be introduced in Section 5, our method of utilizing ethical goal functions aims at actively grounding the implementation of ethics in a transformative socio-technological feedback-loop for which the legislative provides the seed.

4 "Self-Aware" Utility Maximizer

After having commented on the procedure of crafting ethical goal functions, we now describe a class of architectures able to yield controllable utility maximizers that strictly comply with a generic goal function specified by humans. In the following, we explain how a top-down analysis leads to an exemplary technically feasible and minimalistic instance of this class. Note that when we refer to an intelligent system in the following, we specifically mean a system able to independently perform the OODA-loop (Observe, Orient, Decide, Act). One can further decompose the system into four distinct cognitive functions: sensors, orienter, decision maker and actuators according to these four subcomponents

respectively. In a first step, we assume that the utility maximizer cannot be based on a subsymbolic learning paradigm *alone* (such as Deep Learning (DL)), since desirable reactions to all possible situations an intelligent system could encounter in complex real-world environments cannot be learned in reasonable time with finite computational resources. Thus, we postulate in a second step that a certain level of abstraction is required which can be achieved by combining a symbolic reasoning component with a perception exploiting the advantages of learning algorithms resulting in a "hybrid architecture". However, this hybrid intelligent system needs to be as well-equipped with a self-model to face the possible complexity of its internal processes without which the system would be confronted with similar problems caused by the inability to anticipate reactions to all possible internal states. In a third step, we argue that the requirement for a self-awareness capability [1] comprising self-assessment and self-management as well as the ability to provide explanations for actions to human entities appears essential for instance for reasons such as the necessity of constructing solutions in real-time that have not been learned before including sensor management [13], adaptivity in the case of communication to other intelligent systems [14] and for explainability purposes. Apart from this, the view expressed by Thorissón [21] that *"self-modeling is a necessary part of any intelligent being"* which similarly considers the importance of feedback-loops relating the actions of a system to the context of its own internal processes could be a further argument supporting the relevance of self-awareness.

Taking these requirements into account, one feasible instance of the described class of hybrid self-aware utility maximizers could integrate DL algorithms – presently representing relatively accurate Machine Learning models especially in the vision domain – as sensors at the subsymbolic level able to output classification results that can be further processed by the orienter component yielding a symbolic representation of the situation and the internal processes. As decision maker one could envisage a utility-based reasoning/planning (and not learning) process such as e.g. with (partially observable) Markov decision processes (MDP) equipped with the ethical goal function as specified by the legislative, a causal model of the world and of the system itself. The decision maker would map symbolically encoded situations and internal processes to actions maximizing on expected utility with respect to the ethical goal function that are finally executed by the actuators either on the environment or on the system itself. In this framework, explanations could be delivered at the symbolic level. Concerning the input-to-output mappings of the DL sensors, one possibility could be to strive to monitor the related uncertainty by means of self-management which will have to be reflected in the goal function.

5 Socio-Technological Feedback-Loop

Having discussed how a disentanglement of societal responsibilities for the deployment of intelligent systems could be achieved, introduced the notion of an ethical goal function and described the corresponding requirements an intelligent

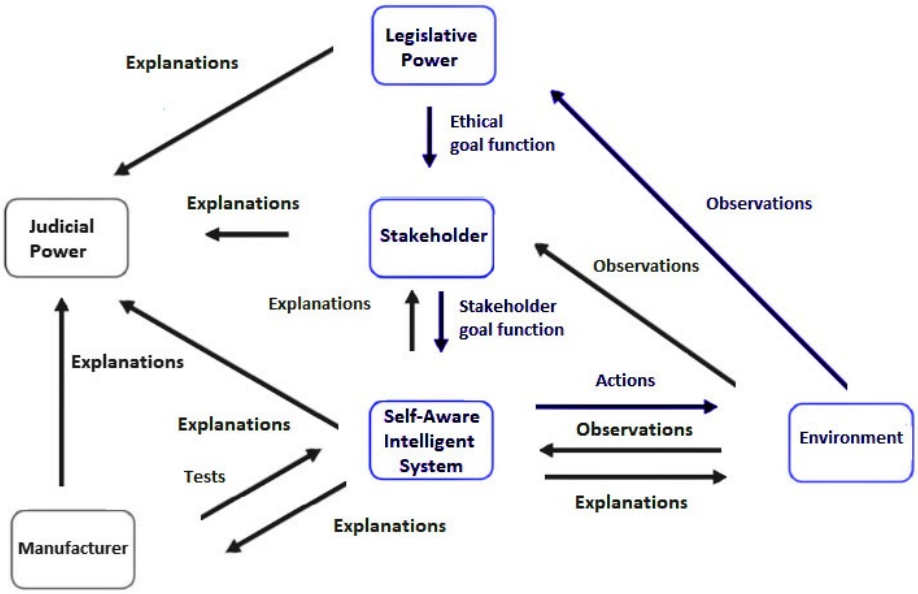

Fig. 1. Simplified illustration and contextualization of a socio-technological feedback-loop (highlighted in blue) implementing the orthogonality-based disentanglement approach for a generic stakeholder domain.

system might need to fulfill in order to comply with such a function, we illustrate and contextualize the composite construction of a consequently resulting socio-technological feedback-loop in Figure 1. At the pre-deployment stage, the manufacturer is responsible for verification and validation practices including the conduct of system tests demonstrating the ability of the intelligent system to adhere to the ethical goal function. At post-deployment stages, the judicial power determines for instance whether the different agents acted in compliance with an ethical goal function given a set of explanations. Concerning the main socio-technological feedback-loop, its key characteristic lies in the fact that it would enable the legislative to dynamically perform revisions of an ethical goal function based on its *quantifiable* impacts on the environment and that it could serve as powerful policy-making tool. Thereby, this feature is paired with the peculiarity that the nature of the environment is not restricted to solely encompass real-world frameworks. More precisely, one could for instance distinguish between three different variations thereof enumerated in an order of potentially increasing speed of formulating/testing hereto related policy-making measures that might be substantiated in an ethical goal function: 1) classical *real-world environments*, 2) specifically crafted and constrained *synthetic environments* and 3) *simulation environments*.

Since the design of an appropriate ethical goal function represents a highly complex task and the necessary time window to collect evidence on its societal

impacts in real-world settings on a large-scale might often represent an undesirable complication, policy experimentation on a small-scale in restricted synthetic environments relating the ethical goal function to specific impacts might represent a complementary measure. However, an even more efficient solution allowing for faster decision-making is the "policy by simulation" approach [23] in which human expert knowledge can be extended by AI systems within simulation environments. In doing so, AI might finally assist humans in developing more ethical AI systems while ultimately enhancing human ethical frameworks by relating the mathematic formulation of an ethical goal function to its direct impacts on the (simulated) environment making possible answers to the crucial question on "what humans *should* want" graspable and beyond that, potentially a direct object of scientific investigation.

Finally, the proposed orthogonality-based disentanglement of responsibilities could provide a new perspective for the AI coordination subtask in AI Safety – the non-trivial issue of making sure that global AI research is dovetailed in such a way that no entity actually implements an unethical and unsafe AGI or ASI – e.g. by offering a starting point for considerations towards an international consensus on the principle of using publicly accessible ethical goal functions that can be easily inspected by the public and international actors. This method might reduce the AI race to the problem-solving ability dimension while at the same time providing incentives for demonstrably ethical and transparent frameworks tightly coupled to an ethical enhancement of partaking societies. Given that the law already represents a public matter, it does thereby not seem to represent an exceedingly disruptive step to advocate for public ethical goal functions.

6 Conclusion and Future Prospects

In a nutshell, the Systems-Engineering oriented approach presented in this paper which we termed "orthogonality-based disentanglement" evinced a technically feasible solution for a responsible deployment of intelligent systems which jointly tackles the control problem and the value alignment problem. We postulated that for this purpose, manufacturers should be responsible for the safety and security of the intelligent systems which they could implement using a utility-based approach with hybrid "self-aware" utility maximizers combining e.g. symbolic reasoning/planning with deep learning sensors. Complementarily, the legislative as representation of the whole society should be responsible for the selection of final goals in the form of human-made, publicly available and quantitatively explicitly specified ethical goal functions (which are not implicitly encoded in an opaque learning model). Additionally, we discussed how a socio-technological feedback-loop stemming from this particular disentanglement might facilitate a dynamical human ethical enhancement supported by AI-driven simulations. Moreover, we briefly explained how the presented framework provides hints on how to solve the AI coordination problem in AI Safety at an international level.

However, certain crucial safety and security challenges remain to be separately addressed and should be taken into consideration in future work. First,

self-improvement within an intelligent system could for instance be implemented by an online learning process or by reconfigurability through run-time adaptivity. While it is reasonable to avoid self-improvement by learning during the deployment of the system in order to limit safety risks, future work will need to examine the possibility of verification methods for self-improvement by reconfigurability at run-time. Second, while the self-awareness functionality facilitates (self-)testing mechanisms, extended research on the controllability of specific test procedures in synthetic testing environments will be required. Third, a turn-off action could be seen as a primitive form of self-management in the context of tasks where the performance of the system superseded human performance. However, the possibility to turn-off the system for security reasons by specified human entities should always be given. Fourth, for the purpose of malevolence prevention, it is important to rigorously consider proactive security measures such as A(G)I Red-Teaming at the post-deployment stage and research on adversarial attacks on the sensors [1, 22] of the self-aware intelligent system. Fifth, a blockchain approach to ensure the security and transparency of the goal functions themselves and all updates on these functions might be recommendable. Crucially, in order to avoid formulations of an ethical goal function with safety-critical side effects for human entities (including implications related to impossibility theorems for consequentialist frameworks [7]), it is recommendable to assign a type of perceiver-dependent and context-sensitive utility to simulations of situations instead of only to the future outcome of actions [3, 2]. In the long-term, we believe that scientific research with the goal to integrate the first-person perspective of society on perceived well-being within an ethical goal function at the core of the presented socio-technological feedback-loop might represent one substantial element needed to promote human flourishing in the most efficient possible way aided by the problem solving ability of AI.

References

1. Aliman, N.M., Kester, L.: Hybrid Strategies Towards Safe "Self-Aware" Superintelligent Systems. In: International Conference on Artificial General Intelligence. pp. 1–11. Springer (2018)
2. Aliman, N.M., Kester, L.: Augmented Utilitarianism. In: International Conference on Artificial General Intelligence. p. to appear. Springer (2019)
3. Aliman, N.M., Kester, L.: Transformative AI Governance and AI-Empowered Ethical Enhancement Through Preemptive Simulations. Delphi - Interdisciplinary review of emerging technologies 2(1), 23–29 (2019)
4. Armstrong, S.: General purpose intelligence: Arguing the orthogonality thesis. Analysis & Metaphysics 12 (2013)
5. Bostrom, N.: The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. Minds and Machines 22(2), 71–85 (2012)
6. Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., et al.: The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. arXiv preprint arXiv:1802.07228 (2018)
7. Eckersley, P.: Impossibility and Uncertainty Theorems in AI Value Alignment (or why your AGI should not have a utility function). CoRR abs/1901.00064 (2018)

8. Elands, P., Huizing, A., Kester, L., Oggero, S., Peeters, M.: Governing Ethical and Effective Behaviour of Intelligent Systems. Military spectator p. to appear (2019)
9. Everitt, T., Lea, G., Hutter, M.: AGI Safety Literature Review. arXiv preprint arXiv:1805.01109 (2018)
10. Goertzel, B.: Infusing advanced AGIs with human-like value systems: Two theses. Journal of Evolution and Technology 26(1), 50–72 (2016)
11. Harris, S.: The moral landscape: How science can determine human values. Simon and Schuster (2011)
12. Hoekstra, R., Breuker, J., Di Bello, M., Boer, A., et al.: The LKIF Core Ontology of Basic Legal Concepts. LOAIT 321, 43–63 (2007)
13. Kester, L., Ditzel, M.: Maximising effectiveness of distributed mobile observation systems in dynamic situations. In: Information Fusion (FUSION), 2014 17th International Conference on. pp. 1–8. IEEE (2014)
14. Kester, L.J.H.M., van Willigen, W.H., Jongh, J.D.: Critical headway estimation under uncertainty and non-ideal communication conditions. 17th International IEEE Conference on Intelligent Transportation Systems (ITSC) pp. 320–327 (2014)
15. Korteling, J.E., Brouwer, A.M., Toet, A.: A neural network framework for cognitive bias. Frontiers in psychology 9 (2018)
16. Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V., Legg, S.: Scalable agent alignment via reward modeling: a research direction. arXiv preprint arXiv:1811.07871 (2018)
17. Pistono, F., Yampolskiy, R.V.: Unethical research: How to create a malevolent artificial intelligence. 25th International Joint Conference on Artificial Intelligence (IJCAI-16). Ethics for Artificial Intelligence Workshop (AI-Ethics-2016) (2016)
18. Van de Poel, I.: Translating values into design requirements. In: Philosophy and engineering: Reflections on practice, principles and process, pp. 253–266. Springer (2013)
19. Russell, S., Dewey, D., Tegmark, M.: Research priorities for robust and beneficial artificial intelligence. AI Magazine 36(4), 105–114 (2015)
20. Sezer, O., Gino, F., Bazerman, M.H.: Ethical blind spots: Explaining unintentional unethical behavior. Current Opinion in Psychology 6, 77–81 (2015)
21. Thórisson, K.R.: A new constructivist AI: from manual methods to self-constructive systems. In: Theoretical Foundations of Artificial General Intelligence, pp. 145–171. Springer (2012)
22. Tomsett, R., Widdicombe, A., Xing, T., Chakraborty, S., Julier, S., Gurram, P., Rao, R., Srivastava, M.: Why the Failure? How Adversarial Examples Can Provide Insights for Interpretable Machine Learning. In: 2018 21st International Conference on Information Fusion (FUSION). pp. 838–845. IEEE (2018)
23. Werkhoven, P., Kester, L., Neerincx, M.: Telling autonomous systems what to do. In: Proceedings of the 36th European Conference on Cognitive Ergonomics. p. 2. ACM (2018)
24. Yudkowsky, E.: The AI Alignment Problem: Why it is Hard, and Where to Start. Symbolic Systems Distinguished Speaker (2016)
25. Yudkowsky, E.: Coherent extrapolated volition. Singularity Institute for Artificial Intelligence (2004)
26. Ziesche, S.: Potential Synergies Between The United Nations Sustainable Development Goals And The Value Loading Problem In Artificial Intelligence. Maldives National Journal of Research 6, 47 (06 2018)