Augmented Utilitarianism for AGI Safety

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

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

Abstract. In the light of ongoing progresses of research on artificial intelligent systems exhibiting a steadily increasing problem-solving ability, the identification of practicable solutions to the value alignment problem in AGI Safety is becoming a matter of urgency. In this context, one preeminent challenge that has been addressed by multiple researchers is the adequate formulation of utility functions or equivalents reliably capturing human ethical conceptions. However, the specification of suitable utility functions harbors the risk of "perverse instantiation" for which no final consensus on responsible proactive countermeasures has been achieved so far. Amidst this background, we propose a novel non-normative socio-technological ethical framework denoted *Augmented Utilitarianism* which directly alleviates the perverse instantiation problem. We elaborate on how augmented by AI and more generally science and technology, it might allow a society to craft and update ethical utility functions while jointly undergoing a dynamical ethical enhancement. Further, we elucidate the need to consider embodied simulations in the design of utility functions for AGIs aligned with human values. Finally, we discuss future prospects regarding the usage of the presented scientifically grounded ethical framework and mention possible challenges.

Keywords: AGI Safety · Utility function · Perverse instantiation · AI alignment · Augmented Utilitarianism

1 Motivation

The problem of unambiguously specifying human goals for advanced AI systems such that these systems once deployed, do not violate the implicitly underlying intended human conceptions by pursuing unforeseen solutions, has been referred to as "literalness" [31,42] or also "perverse instantiation" [10,41] problem. A higher problem solving ability does not necessarily entail the integration of the contextual knowledge required from an advanced AI in order to accurately interpret human ethical conceptions. Therefore, it is of great importance from the perspective of A(G)I Safety and A(G)I Ethics to a priori consider this crucial issue when crafting quantitative utility functions for intelligent systems that would operate based on the human goals these functions encode. Recently, a novel type of such explicitly formulated utility functions denoted *ethical goal*

functions [2,40] has been introduced as critical tool for a society to achieve a meaningful control of autonomous intelligent systems aligned with human ethical values. Departing from this, we show why in order to design ethical goal functions and avoid perverse instantiation scenarios, one needs a novel type of ethical framework for the utility elicitation on whose basis these functions are crafted. We introduce a new to be described socio-technological ethical framework denoted *Augmented Utilitarianism* (which we abbreviate with AU in the following).

While multiple methods have been suggested as moral theory approaches to achieve ethical objective functions for AGIs [16,17] (including classical ethical frameworks like consequentialism or encompassing methods based on uncertain objectives and moral uncertainty [9,15]), most approaches do not provide a fundamental solution to the underlying problem which wrongly appears to be solely of philosophical nature. According to Goertzel [20], *"pithy summaries of complex human values evoke their commonly accepted meanings only within the human cultural context"*. More generally, we argue that in order to craft utility functions that should not lead to a behavior of advanced AI systems violating human ethical intuitions, one has to scientifically consider relevant contextual and embodied information. Moreover, it could be highly valuable to take into account human biases and constraints that obstruct ethical decision-making and attempt to remediate resulting detrimental effects using science and technology. In contrast to the AU approach we will present, most currently known moral theories and classical ethical frameworks considered for advanced AI systems do not integrate these decisive elements and might therefore riskily not exhibit a sufficient safety level with regard to perverse instantiation.

2   Deconstructing Perverse Instantiation

In the following, we enumerate (using the generic notation $< FinalGoal >$: $< PerverseInstantiation >$) a few conceivable perverse instantion scenarios that have been formulated in the past:

1. "Make us smile": "Paralyze human facial musculatures into constant beaming smiles" (example by Bostrom [10])
2. "Make us happy": "Implant electrodes into the pleasure centers of our brains" (example by Bostrom [10])
3. "Making all people happy": "Killing all people [...] as with zero people around all of them are happy" (example by Yampolskiy [42])
4. "Making all people happy": "Forced lobotomies for every man, woman and child [...]" (example by Yampolskiy [42])

From our view, one could extract the following two types of failures out of the specified perverse instantiations: misspecification of final goal criteria and the so called *perspectival fallacy of utility assignment* [2] which will become apparant in our explanation. First, one could argue that already the proposed formulations regarding the criteria of the final goal do not optimally capture the nature of the

intended sense from a scientific perspective which might have finally misguided the AI. While the concept of happiness certainly represents a highly ambiguous construct, modern research in the field of positive psychology [35,39], hedonic psychology [23] and further research areas offers a scientific basis to assess what it means for human entities. For instance, one might come to the conclusion that a highly desirable final goal of humanity for a superintelligence rather represents a concept which is close to the notion of "well-being". In psychology, well-being has been among others described as a construct consisting of five measurable elements: positive emotions, engagement, relationships, meaning and achievement (PERMA) [38]. Another known psychological measure for well-being is subjective well-being [29] (SWB) which is composed of frequent positive affect, infrequent negative affect and life satisfaction [11,13]. In both cases, happiness only represents a subelement of the respective well-being construct. Similarly, as stated by Diener and Bieswas-Diener [14], *"happiness alone is not enough; people need to be happy for the right reasons"*. Coming back to the provided examples for perverse instantiation, in the cases 1, 2 and 4, it is implausible that a pluralistic criteria of well-being like PERMA would have been met.

Second, it is however important to note that even if the final goal would have been specified in a way reflecting psychological insights, a perverse instantiation cannot necessarily be precluded without more ado. By way of illustration, we correspondingly reformulate the example 3 within a new type of perverse instantiation and provide an additional example. We thereby use the term "flourish" to allude to the achievement of a high level of well-being in line with a psychological understanding of the concept as exemplified in the last paragraph.

5. Make all people flourish: Killing all people
6. Make all people flourish: Initiate a secret genocide until the few uninformed people left in future generations all flourish

Despite a suitable final goal, value alignment is not succesful in 5 and 6 because the underlying assignment of utility seems to be based on a detached modus operandi in which the effects of scenarios on the *own* current mental states of the people generating this function are ignored. Thereby, it is assumed that during utility assignment, the involved people are considered as remote observers, while at the same time one inherently takes their perspective while referring to this mapping with the emotionally connoted description of a *perverse* instantiation. This type of detached design of utility functions ignoring i.a. affective and emotional parameters of the own mental state has been described as being subject to the perspectival fallacy of utility assignment [2]. Although most people would currently dislike all provided examples 1–6, the aggregated mental states of their current selves seem not to be reflected within the utility function of the AI which instead considered a synthetic detached measure only related to their future selves or/and future living people. In the next paragraph, we briefly introduce a known problem in population ethics that exhibits similar patterns and which might be of interest for the design of utility functions for advanced AI systems in certain safety-relevant application areas [15].

Population ethics [21] is an issue in philosophy concerning decision-making that potentially leads to populations with varying numbers or/and identities of their members. One interesting element of a population ethics theory is the derived population axiology which represents the total order of different population states according to their ethical desirability. As an example, consider the choice of either perform a policy measure that leads to a population A of ca. 10 billion members and a very high positive welfare or to rather prefer another policy measure leading to a population Z of ca. 10.000 billion members and a much lower only barely acceptable (but still positive) welfare. Intuitively, most people would rank the policy measure leading to population A as higher than the one leading to population Z. However, given the population axiology of total utilitarianism [21], Z might well be ranked higher than A if the number of people multiplied by their welfare is bigger for population Z in comparison to population A. This type of violation of human ethical intuitions when applying total utilitarianism to population ethics has been termed "Repugnant Conclusion" by Derik Parfit [34]. In this context, Arrhenius [3] proved in one of his impossibility theorems that no population axiology[1] can be formulated that concurrently satisfies a certain number of ethical desiderata.

However, as shown by Aliman and Kester [2], this type of impossibility theorem does not apply to population axiologies that take the mental states of those attempting to craft the total orders during utility elicitation into account. Similarly to the perverse instantiation examples 1–6, the application of e.g. total utilitarianism to the described scenario is subject to the perspectival fallacy of utility assignment. As in the case of these perverse instantiations, the fact that most people consider the scenario involving population Z as *repugnant* is not reflected in the utility function which only includes a detached measure of the well-being of future people. In practice, how humans rate the ethical desirability of for instance a policy measure leading to a certain population, is dependent on the effect the mental simulation of the associated scenario has on their corresponding mental states which inherently encode e.g. societal, cultural and temporal information. For instance, from the perspective of a current population Z_0 being similar to population Z both with regard to number of people and welfare level, it might instead be "repugnant" to prefer the policy measure leading to population A [2]. The reason being that the scenario leading from Z_0 to A might have included a dying out or even a genocide. The lack of the required contextual information in consequentialist frameworks (such as utilitarianism) has implications for AIs and AGIs that are implemented in the form of expected utility maximizers mostly operating in a consequentialist fashion.

3   Augmenting Utilitarianism

In the light of the above, it appears vital to refine classical utilitarianism (CU) if one intends to utilize it as basis for utility functions that do not lead to perverse

---

[1] Importantly, this also applies to non-consequentialist frameworks such as deontological ethics [21].

instantiation scenarios. However, as opposed to classical ethical frameworks, AU does not represent a normative theory aimed at specifying what humans *ought to do*. In fact, its necessity arises directly from a technical requirement for the meaningful control of artificial intelligent systems equipped with a utility function. Since the perverse instantiation problem represents a significant constraint to the design of ethical goal functions, a novel tailored ethical framework able to alleviate issues related to both misspecification of final goal criteria and perspectival fallacy of utility assignment emerges as exigency. With this in mind, AU is formulated as a non-normative ethical framework for AGI Safety which can be augmented by the use of science and technology in order to facilitate a dynamical societal process of crafting and updating ethical goal functions. Instead of specifying what an agent ought to do, AU helps to identify what the current society *should want* an (artificial or human) agent to do if this society wants to maximize expected utility. In this connection, utility could ideally represent a generic scientifically grounded (possibly aggregated) measure capturing one or more ethically desirable final goal(s) as defined by society itself. In the following, we describe by what type of components AU could augment CU:

– *Scientific grounding of utility:* According to Jeremy Bentham [7], the founder of CU *"by the principle of utility is meant that principle which approves or disapproves of every action whatsoever according to the tendency it appears to have to augment or diminish the happiness of the party whose interest is in question"*. For AU, one could for instance reformulate the principle of utility by substituting "happiness" with a generic scientific measure for one or more final goal(s). In the context of crafting ethical goal functions, the party whose interest is in question is society. Further, a crucial difference between CU and AU is that in order to assess the tendency an action has to augment or diminish the chosen ethical measure, AU considers more than just the outcome of that action as used in the classical sense, since AU presupposes the *mental-state-dependency* [2] of utility as will be expounded in the next subitem. With this application-oriented view, one could then argue that what society should ideally want an agent to do are actions that are conformable to this modified mental-state-dependent principle of utility. In this paper, we exemplarily consider well-being as reasonable high level final goal candidate which is e.g. already reflected in the UN Sustainable Developmental Goals (SDGs) [44] and is in the spirit of positive computing [12]. Besides SWB [29] and PERMA [38], multiple measures of well-being exist in psychology with focus on different well-being factors. For instance, the concept of objective happiness [23] has been proposed by Kahneman. Well-being has moreover been linked to the hierarchy of needs of Abraham Maslow which he extended to contain self-transcendence at the highest level on top of self-actualization in his later years [26,27,30]. (Recently, related AI research aiming at inducing self-transcendent states for users has been considered by among others Mossbridge and Goertzel [32].) For a review on relevant well-being factors that might be pivotal for a dedicated positive computing, see Calvo and Peters [11].

– *Mental-state-dependency:* As adumbrated in the last section, human ethical evaluation of an outcome of an action is related to their mental states which take into account the simulation that led to this outcome. The mental phenomenon of actively simulating different alternative scenarios (including anticipatory emotions [6]) has been termed conceptual consumption [19] and plays a role in decision-making. Similarly, according to Johnson [22] *"moral deliberation is a process of cognitive conative affective simulation"*. Moreover, it has been shown that for diverse economical and societal contexts, people do not only value the outcome of actions but also assign a well-being relevant *procedural utility* [18,25] to the policy that led to these outcomes. In light of this, AU assigns utility at a higher abstraction level by e.g. considering the underlying state transition (from starting state s over action a to outcome s') instead of the outcome alone as performed in classical consequential frameworks like CU. Furthermore, according to constructionist approaches in neuroscience [5], the brain constructs mental states based on *"sensations from the world, sensations from the body, and prior experience"* [33]. Hence, ethical judgements might vary with respect to multiple parameters encompassing e.g. psychological, biographical, cultural, temporal, social and physiological information. Likewise, the recent Moral Machine experiment studying human ethical conceptions on trolley case scenarios with i.a. autonomous vehicles showed *"substantial cultural variations"* in the exhibited moral preferences [4]. Ethical frameworks for AGI utility functions that disregard the mental-state-dependency may more likely lead to perverse instantiations, since they ignore what we call the *embodied nature of ethical total orders.* In the light of the aforesaid, AU considers perceiver-dependent and context-sensitive utility functions which could e.g. be formulated at the transition level leading to utility functions $U_x(s, a, s')$ for each perceiver x instead of the general $U(s')$ in CU.

– *Debiasing of utility assignment:* One might regard decision utility based on observed choices (as exhibited e.g. in big data [36]) as sufficient utility source for a possible instantiation of AU if one assumes that humans are rational agents that already act as to optimize what increases their well-being. However, utility as measured from this third-person perspective might not capture the actual experienced utility from a first-person perspective due to multiple human cognitive biases [8,24]. Since it is impossible to directly extract the instant utility (the basic building block of experienced utility [24]) of future outcomes to craft ethical goal functions, AU could – in its most basic implementation – rely on predicted utility which represents the belief on the future experienced utility people would assign to a given scenario from a first-person perspective. However, the mental simulations on whose basis predicted utility is extracted are still distorted among others due to the fact that humans fail to accurately predict their appreciation of future scenarios [24]. Therefore, it has been suggested by Aliman and Kester [2] to augment the utility elicitation process by the utilization of technologies like virtual reality and augmented reality within a simulation environment in order to be able to get access to a less biased *artificially simulated future instant utility.* (Thereby, simpler

techniques such as movies or diverse types of immersive storytelling are as well conceivable.) Analogous to the AI-aided policy-by-simulation approach [40], this technique might offer a powerful preemptive tool for AGI Safety in an AU framework. Overall, the experience of possible future world scenarios might improve the quality of utility assignment while having the potential to yield an ethical enhancement for one thing due to the debiased view on the future and secondly, for instance due to beneficial effects that immersive technologies might have on prosocial behavior including certain forms of empathy [12,28]. Interestingly, the experience of individualized and tailored simulations itself might provide an alternative simulation-based solution to the value alignment problem [43].

– *Self-reflexivity:* As opposed to CU, AU is intended as a self-reflexive ethical framework which augments itself. Due to the mental-state-dependency it incorporates and the associated embodied nature of ethical total orders, it might even be necessary to craft new ethical goal functions within a so-called socio-technological feedback-loop [2]. In doing so, ongoing technological progresses might help to augment the debiasing of utility assignment while novel scientific insights might facilitate to filter out the most sophisticated measure for the ethically desired form of utility given the current state of society. Advances in A(G)I development itself leading to a higher problem solving ability might further boost AU with an improved predictability of future outcomes leading to more precise ethical goal functions. Given its generic nature, what humans should want an agent to do might thereby vary qualitatively in an AU framework, since quantitatively specifiable observations at specific time steps within a socio-technological feedback-loop might even lead society to modify the desired final goal candidate(s) making it possible to ameliorate the framework as time goes by.

Table 1. Decision-making focuses within different possible ethical frameworks for AGI Safety. "S&T" denotes a foreseen augmentation of the ethical decision making process by science and technology including A(G)I itself. By "experiencer", we refer to the entities in society performing the ethical evaluation via the experience of simulations (in a mental mode only or augmented).

Ethics framework/Focus	Agent	Action	Outcome	Experiencer	S&T
Virtue ethics	x				
Deontological ethics		x			
Consequentialist ethics (e.g. CU)			x		
AU	**x**	**x**	**x**	**x**	**x**

– *Amalgamation of diverse perspectives*: Finally, we postulate that AU[2], despite its intrinsically different motivation as a socio-technological ethical framework for AGI Safety and its non-normative nature, can be nevertheless understood as allowing a coalescence of diverse theoretical perspectives that have been historically assigned to normative ethical frameworks. To sum up and contextualize the experiencer-based AU, Table 1 provides an overview on the different decision-making focuses used in relevant known ethical frameworks that might be seen as candidates for AGI Safety.

4   Conclusion and Future Prospects

In a nutshell, we proposed AU as a novel non-normative socio-technological ethical framework grounded in science which is conceived for the purpose of crafting societal ethical goal functions for AGI Safety. While CU and other classical ethical frameworks if used for AGI utility functions might engender the perverse instantiation problem, AU directly tackles this issue. AU augments CU by the following main elements: scientific grounding of utility, mental-state-dependency, debiasing of utility assignment using technology, self-reflexivity and amalgamation of diverse perspectives. Thereby, AU facilitates the explicit formulation of perceiver-dependent and context-sensitive utility functions (e.g. of the form $U_x(s, a, s')$ instead of $U(s')$ as performed in CU) for an aggregation at the societal level. These *human-crafted* ethical goal functions should be made publicly available within a white-box setting e.g. for reasons of transparency, AI coordination, disentanglement of responsibilities for AI governance and law enforcement [2] (which differs from using utility functions implicitly learned by AI agents or AIs learning moral conceptions from data such as e.g. in [36]). Besides being able to contribute to the meaningful control of intelligent systems, AU could also be utilizable for human agents in the policy-making domain. Overall, we agree with Goertzel [20] that the perverse instantiation problem seems rather not to represent *"a general point about machine intelligence or superintelligence"*.

One of the main future challenges for the realization of AU could be the circumstance that one can only strive to approximate ethical goal functions, since a full utility elicitation on all possible future scenarios is obviously not feasible. However, already an approximation process within a socio-technological feedback-loop could lead to an ethical enhancement at a societal level. Besides that, in order to achieve safe run-time adaptive artificial intelligent systems reliably complying with ethical goal functions, a "self-awareness" functionality might be required [1,40]. Moreover, the security of the utility function itself is essential, due to the possibility of its modification by malevolent actors during the deployment phase. Finally, proactive AGI Safety research [1] on *ethical adversarial examples* – a conceivable type of integrity attacks on the AGI sensors having ethical consequences might be important to study in future work to complement the use of safe utility functions.

---

[2] AU is not be to confused with agent-relative consequentialism which is a normative agent-based framework, does not foresee a grounding in science and seems to assume a "pretheoretical grasp" [37] of its "better-than-relative-to" relation.

References

1. Aliman, N.-M., Kester, L.: Hybrid strategies towards safe self-aware superintelligent systems. In: Iklé, M., Franz, A., Rzepka, R., Goertzel, B. (eds.) AGI 2018. LNCS (LNAI), vol. 10999, pp. 1–11. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-97676-1_1
2. Aliman, N.M., Kester, L.: Transformative AI governance and AI-empowered ethical enhancement through preemptive simulations. Delphi Interdisc. Rev. Emerg. Technol. **2**(1), 23–29 (2019)
3. Arrhenius, G.: An impossibility theorem for welfarist axiologies. Econ. Philos. **16**(2), 247–266 (2000)
4. Awad, E., et al.: The moral machine experiment. Nature **563**(7729), 59 (2018)
5. Barrett, L.F.: The theory of constructed emotion: an active inference account of interoception and categorization. Soc. Cogn. Affect. Neurosci. **12**(1), 1–23 (2017)
6. Baucells, M., Bellezza, S.: Temporal profiles of instant utility during anticipation, event, and recall. Manag. Sci. **63**(3), 729–748 (2016)
7. Bentham, J.: An Introduction to the Principles of Morals and Legislation. Dover Publications, Mineola (1780)
8. Berridge, K.C., O'Doherty, J.P.: From experienced utility to decision utility. In: Neuroeconomics, pp. 335–351. Elsevier (2014)
9. Bogosian, K.: Implementation of moral uncertainty in intelligent machines. Mind. Mach. **27**(4), 591–608 (2017)
10. Bostrom, N.: Superintelligence: Paths, Dangers, Strategies, 1st edn. Oxford University Press Inc., New York (2014)
11. Busseri, M.A., Sadava, S.W.: A review of the tripartite structure of subjective well-being: implications for conceptualization, operationalization, analysis, and synthesis. Pers. Soc. Psychol. Rev. **15**(3), 290–314 (2011)
12. Calvo, R.A., Peters, D.: Positive Computing: Technology for Wellbeing and Human Potential. MIT Press, Cambridge (2014)
13. Diener, E.: Subjective well-being: the science of happiness and a proposal for a national index. Am. Psychol. **55**(1), 34 (2000)
14. Diener, E., Biswas-Diener, R.: Happiness: Unlocking the Mysteries of Psychological Wealth. Wiley, New York (2011)
15. 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)
16. Everitt, T.: Towards safe artificial general intelligence. Ph.D. thesis, Australian National University (2018)
17. Everitt, T., Lea, G., Hutter, M.: AGI safety literature review. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, pp. 5441–5449. International Joint Conferences on Artificial Intelligence Organization, July 2018. https://doi.org/10.24963/ijcai.2018/768
18. Frey, B.S., Stutzer, A.: Beyond Bentham-measuring procedural utility (2001)
19. Gilbert, D.T., Wilson, T.D.: Prospection: experiencing the future. Science **317**(5843), 1351–1354 (2007)
20. Goertzel, B.: Superintelligence: fears, promises and potentials. J. Evol. Technol. **24**(2), 55–87 (2015)

21. Greaves, H.: Population axiology. Philos. Compass **12**(11), e12442 (2017)
22. Johnson, M.: Moral Imagination: Implications of Cognitive Science for Ethics. University of Chicago Press, Chicago (1994)
23. Kahneman, D., Diener, E., Schwarz, N.: Well-Being: Foundations of Hedonic Psychology. Russell Sage Foundation, New York (1999)
24. Kahneman, D., Wakker, P.P., Sarin, R.: Back to Bentham? explorations of experienced utility. Q. J. Econ. **112**(2), 375–406 (1997)
25. Kaminitz, S.C.: Contemporary procedural utility and Hume's early idea of utility. J. Happiness Stud. **20**, 1–14 (2019)
26. Kaufman, S.B.: Self-actualizing people in the 21st century: integration with contemporary theory and research on personality and well-being. J. Humanist. Psychol. 0022167818809187 (2018). https://doi.org/10.1177/0022167818809187
27. Koltko-Rivera, M.E.: Rediscovering the later version of Maslow's hierarchy of needs: self-transcendence and opportunities for theory, research, and unification. Rev. Gen. Psychol. **10**(4), 302–317 (2006)
28. van Loon, A., Bailenson, J., Zaki, J., Bostick, J., Willer, R.: Virtual reality perspective-taking increases cognitive empathy for specific others. PloS ONE **13**(8), e0202442 (2018)
29. Lyubomirsky, S.: Why are some people happier than others? The role of cognitive and motivational processes in well-being. Am. Psychol. **56**(3), 239 (2001)
30. Maslow, A.H.: The Farther Reaches of Human Nature. Viking Press, New York (1971)
31. Meuhlhauser, L., Helm, L.: Intelligence explosion and machine ethics. In: Singularity Hypotheses: A Scientific and Philosophical Assessment, pp. 101–126 (2012)
32. Mossbridge, J., et al.: Emotionally-sensitive AI-driven android interactions improve social welfare through helping people access self-transcendent states. In: AI for Social Good Workshop at Neural Information Processing Systems 2018 Conference (2018)
33. Oosterwijk, S., Lindquist, K.A., Anderson, E., Dautoff, R., Moriguchi, Y., Barrett, L.F.: States of mind: emotions, body feelings, and thoughts share distributed neural networks. NeuroImage **62**(3), 2110–2128 (2012)
34. Parfit, D.: Reasons and Persons. Oxford University Press, Oxford (1984)
35. Peterson, C.: A Primer in Positive Psychology. Oxford University Press, Oxford (2006)
36. Rafal, R., Kenji, A.: Toward artificial ethical learners that could also teach you how to be a moral man. In: IJCAI 2015 Workshop on Cognitive Knowledge Acquisition and Applications (Cognitum 2015). IJCAI (2015)
37. Schroeder, M.: Teleology, agent-relative value, and 'good'. Ethics **117**(2), 265–295 (2007)
38. Seligman, M.E.: Flourish: A Visionary New Understanding of Happiness and Well-Being. Simon and Schuster, New York (2012)
39. Seligman, M.E.P., Csikszentmihalyi, M.: positive psychology: an introduction. In: Csikszentmihalyi, M. (ed.) Flow and the Foundations of Positive Psychology, pp. 279–298. Springer, Dordrecht (2014). https://doi.org/10.1007/978-94-017-9088-8_18
40. 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)
41. Yampolskiy, R.V.: Utility function security in artificially intelligent agents. J. Exp. Theor. Artif. Intell. **26**(3), 373–389 (2014)

42. Yampolskiy, R.V.: Artificial Superintelligence: A Futuristic Approach. Chapman and Hall/CRC, Boca Raton (2015)
43. Yampolskiy, R.V.: Personal universes: a solution to the multi-agent value alignment problem. arXiv preprint arXiv:1901.01851 (2019)
44. Ziesche, S.: Potential synergies between the united nations sustainable development goals and the value loading problem in artificial intelligence. Maldives Nat. J. Res. **6**, 47 (2018)