Transformative AI Governance and AI-Empowered Ethical Enhancement Through Preemptive Simulations

*Nadisha-Marie Aliman and Leon Kester**

As the problem-solving ability of AI increased significantly during this decade, its scope of application has been extended to various areas including ethically relevant fields such as the development of autonomous systems. In this contribution, we evince why for the purpose of effective AI governance, humans have to quantitatively specify their ethical conceptions within a consequential framework. Thereby, the implementation of advanced AI systems not only forces society to provide machine-understandable ethical goal functions, but it also simultaneously facilitates a new transformative socio-technological feedback-loop with the potential for a dynamic ethical enhancement at the societal level. Furthermore, we exemplify why a common objection to consequentialism related to impossibility theorems does not represent an argument against the feasibility of such a consequential framework with ethical goal functions despite the soundness of these theorems. Finally, we elaborate on how AI (and broader science and technology) might equip humans with a novel particularly powerful preemptive tool within a socio-technological feedback-loop: the ability to get access to a simulated first-person experience of future states of the world and the estimation of the related – as we term it – artificially simulated future instant utility.

I. Consequential AI Governance Strategy

The pertinent progress in research on intelligent systems exhibiting a higher and higher problem-solving ability confronts society with the need to select appropriate AI governance strategies in order to identify the required legal and ethical framework. In this context, one could distinguish four main conceivable candidate clusters of strategies to govern AI: 1) *prohibitive*, 2) *self-regulative*, 3) *deontological* and 4) *consequential* approaches. In the following, we will explain how for different social and technical Systems-Engineering oriented reasons, solution 4) represents the only recommendable AI governance strategy from this pool. First, the prohibitive strategy 1) aiming at fundamentally restricting or even banning research on advanced AI systems can be classified as an approach with a highly unlikely practicability given the incentives for technological progress and is thus not further considered. Second, method 2) foresees self-regulative mechanisms which might be inherent to the market or to the specific architectural design of the intelligent systems and might account for the emergence of a certain stability after the deployment of these systems. However, since the AI landscape is highly heterogeneous, society could not rely on the conception that safe, secure and ethical designs are necessarily preeminent on the market and would moreover face confusing entanglements within the assignment of responsibilities to specific users, manufacturers, operators or legislators. Since it therefore appears unfeasible to ensure a sufficient level of controllability within a deployment of intelligent systems in accordance with strategy 2), it is not further considered in this analysis. At first sight, the remaining feasible strategies seem to be the de-

DOI: 10.21552/delphi/2019/1/6

* Nadisha-Marie Aliman, M.Sc., PhD candidate at Utrecht University Department of Information and Computing Sciences. For correspondence: nadishamarie.aliman@gmail.com.
Dr. Leon Kester, Senior Research Scientist on ethical intelligent systems, TNO Netherlands.

The authors would like to thank Peter Eckersley for a discussion on ethical utility functions and impossibility theorems.

ontological method 3) whose goal is to embed ethical values in AI systems via deontological rules and the consequential approach 4) for which ethical values have to be *quantitatively* encoded into machine-readable mathematical objective functions. Generally, it can be assumed that it is in the interest of a democratic society that the ethical framework utilised for intelligent systems is determined by society itself or a suitable representation of society such as the legislative power. In this context, a transparent disentanglement of responsibilities ensuring that the systems act in accordance with ethical and legal frameworks as specified by the legislative power and facilitating the attribution of responsibilities by the judicial power would be made possible. On a technical level, one would thereby need an approach able to actually practically realise the necessary disentanglement of the *what* and the *how*. More precisely, it has to be a technically feasible method within which the final (ethical) goals of an intelligent system (the *what*) and its problem-solving ability (the *how*) are orthogonal[1] to each other. We denote this type of technical Systems-Engineering oriented solution for a responsible governance of intelligent systems *orthogonality-based disentanglement of responsibilities*. In the case of both methods 3) and 4), legislators could theoretically be responsible for the ethical framework and the manufacturers for the technical implementation of the intelligent systems including their safety and security. However, in the next paragraph, we will briefly enumerate a number of rationales that exemplify why the deontological method 3) cannot be seen as a possible instantiation of that disentanglement procedure, leaving the consequential method 4) as the only realistic AI governance strategy.

First the attempt of method 3) to try to formulate deontological rules for every situation an intelligent system might encounter in a complex real-world environment is technically impracticable (it leads to a 'state-action space explosion').[2] Conversely, for the consequential strategy 4), there exist corresponding Systems Engineering oriented techniques on how to implement run-time adaptive models equipped with a so-called 'self-awareness' functionality (self-management, self-assessment and the ability to provide explanations)[3] that would not face such problems. Second, since law is formulated in natural language which is intrinsically ambiguous on multiple linguistic levels, either an intelligent system implemented in accordance with method 3) will have to extract meaning out of this text material using fault-prone Natural Language Processing techniques or the developers might make use of ontologies encoding law which would however require them to first interpret law, which would in turn violate the idea of disentangling responsibilities. Using approach 4), one could circumvent these drawbacks by crafting unambiguous mathematical functions formulated by (a representation of) society. In the following, these objective functions that should encode the ethical and legal framework are referred to as *ethical goal functions*. Third, legal frameworks often leave trade-offs and dilemmas open which the deontological approach cannot directly solve, a problem which a consequential system would not encounter. Fourth, an update of laws in the deontological case will require every manufacturer to costly modify the built-in ethical framework, while the consequential solution would only require a centralised update of an ethical goal function. Fifth, the mathematically defined nature of approach 4) opens up new possibilities for a dynamical AI-empowered ethical enhancement of society and might – with an ethical goal function as its core – generate a beneficial socio-technological feedback-loop (as will be introduced in Section II) which a deontological approach cannot afford. Therefore, the consequential strategy can be regarded as the only both feasible and desirable instantiation of the *orthogonality-based disentanglement of responsibilities* required if society is willing to realise efficient AI governance measures.

II. Dynamical Ethical Enhancement through a Socio-Technological Feedback-Loop

The realisation of a consequential approach to AI governance utilising ethical goal functions should be

1 Nick Bostrom, 'The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents' (2012) 22 Minds and Machines 2, 71-85

2 Peter Werkhoven et al, 'Telling Autonomous Systems What To Do' (36th European Conference on Cognitive Ergonomics, Utrecht, September 2018)

3 Nadisha-Marie Aliman and Leon Kester, 'Hybrid Strategies Towards Safe 'Self-Aware' Superintelligent Systems' (International Conference on Artificial General Intelligence, Prague, August 2018)

considered as a dynamical process in which these functions are steadily reviewed and updated. Given a domain, the legislative could provide an ethical goal function to a given stakeholder. This ethical goal function would quantitatively specify the utility of every outcome of actions an intelligent system might select in that domain while the stakeholder operates a system which would have to perform actions maximising the expected utility given that (potentially customised) function. During the deployment of the intelligent system, the system provides explanations for its actions to the stakeholder while the legislative as well as policy-makers have the possibility to collect observations on the environment within which the intelligent systems carry out actions. Based on this analysis considering quantifiable ethical impacts, a new scientifically grounded update of the ethical goal function can be undertaken. In a next step, the legislative provides the new updated goal function to the stakeholder by which the loop starts anew. (Thereby, the role of the manufacturer is to provide sufficient security and safety testing measures *before* the deployment of an intelligent system. Finally, *after* the deployment of the system, the judicial power is able to adequately assign responsibilities to participating entities given their explanations.) We call this loop within which society achieves an ethical enhancement through the use of technology the *socio-technological feedback-loop*. Importantly, this feedback-loop is not restricted to an implementation within real-world environments, since an AI-aided technique called 'policy by simulation'[4] enables the generation of what-if scenarios via simulations in a much more time-efficient, cost-efficient and safer way. As a result, policy-makers can perform policy experimentation with different goal functions in simulation environments which facilitates the choice of appropriate safe ethical goal functions. Moreover, since the ethical goal functions represent a type of encoding of ethics, AI might enable society to implement more ethical AI systems and by doing this ultimately enhance human ethical thinking.

From the perspective of AI safety, this socio-technological feedback-loop might immanently solve the 'control problem' and the 'value alignment problem' – with the former being the task on how to build advanced AI systems that do not harm humans and the latter addressing how to implement AI that is aligned with human values. Likewise, it is cogitable that if multiple societies at an international level opt for this type of governance solution with ethical goal functions, which, as in the case of classical laws will have to be made publicly accessible, will promote transparency and safety of global AI research while fostering the efficient development of more ethical frameworks. Achieving an international consensus on using this strategy might thereby additionally represent a solution to the 'AI coordination problem' which is 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 advanced AI in the first place. However, the success of initiating an approach based on ethical goal functions will be crucially dependent on the quality of the procedure of utility assignment consisting of a mapping of utility values to states of the world as required to be performed by (a representation of) society. In the next section, we will address a common apparently weighty objection against the feasibility of such a clear assignment within consequentialist frameworks and explain why it does not affect the design of ethical goal functions for artificial intelligent systems. What is more, we point out a fundamental misconception underlying that objection. Finally, in a further section, we elaborate on a conceivable possibly futuristic seeming research direction that might contribute to obtain utility assignments of an improved quality by allowing humans to in a sense experience future well-being in the present.

III. Implications of Impossibility Theorems for AI Governance and Ethical Enhancement

Possible areas of application for intelligent systems encompass ethically relevant contexts within which the decision-making process might directly affect the well-being of currently living people or populations of people that might exist in the future.[5] For this reason, it is of critical importance to make sure that ethical goal functions are able to safely encode desirable

4 (n 2)

5 Peter Eckersley, 'Impossibility and Uncertainty Theorems in AI Value Alignment (or why your AGI should not have a utility function' (AAAI's Workshop on Artificial Intelligence Safety, Honolulu, January 2019)

conceptions on population ethics which are not in conflict with those of the society that crafted it. Population ethics[6] is an area of philosophy addressing ethical issues concerning 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 ordering of different population states according to their ethical desirability. By way of illustration, consider the simplified example of comparing a *population A* of ca. 10 billion members and a very high positive welfare with a *population Z* of ca. 10.000 billion members and a much lower only barely acceptable but still positive welfare. At first sight, it seems that population A should be ranked higher than population Z, since it appears to be the ethically preferable population state of both if one had the choice. However, the naïve application of total utilitarianism to this example leads to the circumstance that 'any loss in the quality of lives in a population can be compensated for by a sufficient gain in the quantity of a population'[7] which might potentially lead to the solution that population Z should be preferred to population A. This would be the case if the area below the welfare curve – here simply representing the number of people multiplied by their welfare – is bigger for population Z in comparison to population A. This non-intuitive and potentially unethical type of result when applying total utilitarianism to population ethics has been termed a 'Repugnant Conclusion' by Derik Parfit.[8]

Diverse mathematical and philosophical approaches to avoid this 'Repugnant Conclusion' have been studied, but led to the insight that reasonable approaches able to avoid this conclusion entail one or more comparable unethically seeming conclusion(s) as shown by Arrhenius[9] in one of his impossibility theorems. More precisely, he proved that *no* welfarist population axiology can concurrently satisfy a certain number of required ethical desiderata.[10] This means that a complete ranking of states of populations (mathematically corresponding to a *total order* of these populations) according to their ethical desirability is not possible.[11] Prima facie, this circumstance might pose a potential obstacle to the implementation of intelligent systems equipped with an *ethical* goal function assigning utility to states of the world for instance related to the well-being of people, since it seems as if this utility assignment could not be performed in the first place without inherently leading to one or more *unethical* conclusion(s). However, we will elaborate on how despite their soundness, impossibility theorems asserting the impossibility of an unambiguously ethical welfarist population axiology (and thus the impossibility of a corresponding ethical total order over possible population states) do not represent a valid argument against the viability of crafting ethical goal functions in order to achieve ethical intelligent systems.

In the example comparing population A to population Z, it was assumed that the utilitarian observer(s) performing the assignment of utility to each of these population states would allot the higher utility to the population state for which the area below the welfare curve is bigger. Thereby, the utilitarian observer(s) would assume a third-person perspective, since a remote measure of the welfare of people within the populations is considered. However, by doing this, a detachment from any *own* hedonic utility is actually taking place. We designate this detachment as the *perspectival fallacy of utility assignment*. We argue that in fact, utilitarian decision-making should not be necessarily regarded as a remote, detached and passive endeavour, but could instead be implemented as an active task based on the *own* experienced utility (as perceived from a first-person perspective) that arises in real-time while mentally evaluating and thereby simulating the different alternative scenarios. For it is eg known that 'anticipatory emotions arise in reaction to mental discrete images of the outcome of a decision'[12] and that this mental simulation phenomenon termed *conceptual consumption*[13] provides a basis for decision-making. Moreover, to consider the thereby experienced utility in this immediate hedonic sense is much closer to the original idea

6 Hilary Greaves, 'Population Axiology' (2017) 12 Philosophy Compass 11

7 Gustaf Arrhenius, 'An Impossibility Theorem for Welfarist Axiologies' (2000) 16 Economics & Philosophy 2, 247-266

8 Derek Parfit, *Reasons and Persons* (OUP Oxford, 1984)

9 (n 7)

10 (n 6)

11 Note, that this finding does not only apply to consequential frameworks, since every moral theory needs a population axiology which is why eg deontological analogues for impossibility theorems are similarly conceivable. See, (n 6)

12 M Baucells and S Bellezza 'Temporal Profiles of Instant Utility during Anticipation and Recall' (2014) Management Science

13 Daniel T Gilbert and Timothy D Wilson, 'Prospection: Experiencing the Future' (2007) 317 Science 5843, 1351-1354

of 'utility' as introduced by Jeremy Bentham.[14] Therefore, we argue that a society willing to perform a utility assignment with the goal to achieve a population axiology, could rate different populations states according to the aggregated experienced utility that the simulation of these states generates in the minds of the members of this society.

When further considering this type of utility elicitation, it becomes clear that the utility that a utilitarian would assign to an outcome would be dependent on its mental state which might eg inherently encode individual psychological, temporal, biographical, social and cultural information. In the case of a utilitarian society, the overall resulting utility would encode an aggregation of the mental states of all its members. In the following, we refer to this general dependence on mental states as the *state-dependence of population axiology*. Due to this dependence, it is cogitable that *different mental states could potentially lead to different population axiologies* ie varying mental states could lead to varying total orders over population states. Now reconsidering the impossibility theorem of Arrhenius stating that no welfarist axiology can simultaneously satisfy a number of required ethical desiderata, it becomes however clear that he actually examined the possibility of the *one* single absolute context-independent and state-independent axiology given a population ethics framework. Therefore, what was proven is only that *no* single *state-independent axiology* can simultaneously satisfy a number of ethical desiderata. This lets the eventuality untouched that a utility assignment considering the first-person perspective of a society performing that assignment might be able to lead to a *state-dependent* axiology which could simultaneously satisfy a number of ethical desiderata. More precisely, it might still be possible that a state-dependent total order of population states would be achievable without entailing any unethically seeming conclusion.

For illustrative purposes, one could reconsider the example with population A and population Z, but now considering utility assignments based on own experienced utility. Further, we assume that both populations are future populations that could result out of a policy making measure that the society which performs the utility assignment might take or not take. In today's society, it appears intuitively ethical to prefer population A, because for most people, mentally simulating the future population A seems to have a higher positive intensity than the case with the future population Z. This is well reflected in the emotionally connoted use of the term 'Repugnant Conclusion' in the case population Z would be preferred instead. However, one could conversely for instance imagine that the current society performing the utility assignment is similar to population Z both with regard to the number of persons and their welfare. Supposing that this society would like to perform a utility assignment for a policy measure that should either transform society towards population A in the future or rather keep it in a similar form with the same number of people and the same welfare, it is easily comprehensible that a different conclusion might arise. Namely, it is possible that this society might perceive the option with population A as a dying out or even as a genocide and would, despite the higher welfare level, assign higher utilities to population Z due to the negatively valanced mental simulation of this scenario. This new total order placing population Z before population A would however appear natural to most people. This circumstance can be explained by the introduced state-dependence of population axiology. To sum up, coming back to the realisation of ethical intelligent systems via ethical goal functions, we showed that an impossibility theorem for consequentialist frameworks does not represent a valid argument against the possibility for a society to actually craft these ethical goal functions which are inherently of state-dependent nature. In that respect, a dynamical update of ethical goal functions as society evolves towards different states along the time axis within a socio-technological feedback-loop might even be *necessary* since different total orders of population states could be suitable for different distinct states that society might reach as time goes by.

IV. Experiencing Future Well-Being in the Present

As described in the last section in the context of the perspectival fallacy of utility assignment, it is expedient to consider utility as being grounded in hedonic experience from a first-person perspective. Ad-

14 Jeremy Bentham, *The Collected Works of Jeremy Bentham: An Introduction to the Principles of Morals and Legislation* (Clarendon Press, 1996)

mittedly, so far, we did not concretise how to objectively measure this experienced utility which might however be crucial for the process of crafting ethical goal functions. For one thing, one might question the scientific measurability of hedonic experience in the first place. Secondly, one might assume that experienced utility can if measurable, be indirectly inferred from a third-person perspective via observed choices of individuals from which the so-called decision utility – potentially already reflecting hedonic experience – is often extracted in economics. However, as shown by Kahneman in multiple studies[15] experienced utility can indeed be measured and used for interpersonal comparisons. Moreover, he demonstrated that decision utility is not necessarily congruent with experienced utility due to multiple human cognitive biases. Thus, in the following, we presuppose that experienced utility is objectively measurable and represents a more realistic model of hedonic experience which is directly linked to human well-being/happiness. While other approaches considering a first-person perspective on experienced well-being are certainly possible, this contribution exemplarily focuses on the assumption made by Kahneman according to which experienced utility can be measured via its basic building block termed *instant utility*. He describes instant utility as being 'a measure of hedonic and affective experience, which can be derived from immediate reports of current subjective experience or from physiological indices'.[16] For subjective experiences spanning over a certain time slot, Kahneman introduces the notion of total utility which is constructed from temporal profiles of instant utility. (More precisely, he defines it as being the temporal integral of instant utility.) Further, he assumes that objective happiness represents the average utility given a certain period of time.[17] Thus, the consideration of instant utility can be seen as a bottom-up approach to well-being/happiness.

Having introduced what could be the basic measure for the experienced utility assignment procedure, it is important to note that instant utility would capture the immediate hedonic experience *while* the outcome of a certain decision is taking place. However, one has to craft ethical goal functions *before* the outcomes of actions performed by the intelligent system take place. This requirement seems impossible to fulfill. The only practicable approximation seems to be a *predicted utility* representing our belief on the experienced utility we might experience from a future outcome. With other words, individuals might envisage a future scenario and assign utility according to the effect this mental simulation has on them. However, experiments led to the conclusion that predicted utility is subject to diverse considerable cognitive biases and often crucially differs from instant utility. For instance, it has been shown that people exhibit a 'limited understanding and ability to predict their own enjoyment of goods and activities'.[18] Since this circumstance might lead to ethical goal functions that do not maximise on the actually desired objective of happiness/well-being and this might even lead to safety issues, we argue that it is important to complement the utilisation of predicted utility with sophisticated proactive measures.

Given current technological advancements including the possibility to perform AI-aided preemptive techniques for policy-making like 'policy by simulation' (as mentioned in chapter II), we argue that it might similarly be possible to approximate the instant utility of future outcomes more accurately by means of simulation environments. In the future, such preemptive policy experimentation procedures could allow society (or a representation thereof) to directly experience scenarios leading to future states of the world as computed by AI systems for instance within a simulated virtual reality or augmented reality environment. By doing this, society might literally be able to experience (an approximation of) future well-being in the present. During this simulated future experience, one might use respective methods to measure instant utility (and the total utility computed therefrom) in real-time. We call this type of experienced utility *artificially simulated future instant utility*. Depending on the quality of the simulations, it is thinkable that this *artificially simulated future instant utility* would represent a much better approximation of the true instant utility that the outcome

15 Daniel Kahneman et al, 'Back to Bentham? Explorations of Experienced Utility' (1997) 112 The Quarterly Journal of Economics 2, 375-406; Daniel Kahneman, Edward Diener, and Norbert Schwarz (eds) *Well-being: Foundations of hedonic psychology* (Russell Sage Foundation, 1999)

16 Daniel Kahneman et al, 'Back to Bentham? Explorations of Experienced Utility' (1997) 112 The Quarterly Journal of Economics 2, 375-406

17 Daniel Kahneman, Edward Diener, and Norbert Schwarz (eds) *Well-being: Foundations of Hedonic Psychology* (Russell Sage Foundation, 1999)

18 Daniel Kahneman et al, 'Back to Bentham? Explorations of Experienced Utility' (1997) 112 The Quarterly Journal of Economics 2, 375-406

would elicit than it would be the case for the predicted utility. While predicted utility is among others mainly based on a mental simulation distorted by human biases which could lead to safety-critical errors, realistic simulation environments might provide a more concise estimation and thus a better and safer assessment on how different outcomes of actions that intelligent systems might take finally relate to human well-being.

V. Conclusion

If it holds that objective happiness represents a suitable bottom-up approach to well-being/happiness, then an ideal strategy to promote human well-being, would be to implement ethical intelligent systems able to maximise on the aggregated simulated future instant utility (ie the correspondingly aggregated total utility) that a society experienced during the preemptive simulations of states of the world. In this ideal world, ethical goal functions would serve exactly this purpose. However, besides the fact that AI models are not omniscient and might not be able to always yield reliable predictions of future world states, it is obvious that a utility assignment by society on all possible outcomes is not feasible. Therefore, this full utility assignment reflecting the aggregated artificially simulated future instant utility of society with regard to all states of the world can only be complemented and approximated by AI models via cardinal ethical goal functions with multiple parameters. (Note that one could also consider to conceptually incorporate parameters derived from top-down approaches to well-being such as eg the PERMA model of positive psychology[19] which similarly considers a first-person perspective on experienced well-being.) However, we presume that already a dynamic update of such approximate ethical goal functions might offer a huge potential to promote human well-being using intelligent systems. Overall, it can therefore be summarised that the presented consequential AI governance approach would not only be able to address fundamental global issues such as the AI value alignment problem, but it would also facilitate a transformative socio-technological feedback-loop with unseen opportunities for the ethical enhancement of humans and their pursuit of well-being. Importantly, we further showed that despite the soundness of impossibility theorems for classical consequentialist frameworks, these theorems do not entail the impossibility of the proposed transformative AI governance strategy which is based on state-dependent ethical goal functions.

19 Martin EP Seligman, *Flourish: A Visionary New Understanding of Happiness and Well-Being* (Simon and Schuster, 2012)