

Requisite Variety in *Ethical* Utility Functions for AI Value Alignment

Nadisha-Marie Aliman¹, Leon Kester²

¹ Utrecht University, Utrecht, Netherlands

² TNO Netherlands, The Hague, Netherlands

nadishamarie.aliman@gmail.com

Abstract

Being a complex subject of major importance in AI Safety research, value alignment has been studied from various perspectives in the last years. However, no final consensus on the design of ethical utility functions facilitating AI value alignment has been achieved yet. Given the urgency to identify systematic solutions, we postulate that it might be useful to start with the simple fact that for the utility function of an AI not to violate human ethical intuitions, it trivially has to be a *model* of these intuitions and reflect their *variety* – whereby the most accurate models pertaining to human entities being biological organisms equipped with a brain constructing concepts like moral judgements, are *scientific* models. Thus, in order to better assess the variety of human morality, we perform a trans-disciplinary analysis applying a security mindset to the issue and summarizing variety-relevant background knowledge from neuroscience and psychology. We complement this information by linking it to *augmented utilitarianism* as a suitable ethical framework. Based on that, we propose first practical guidelines for the design of approximate *ethical goal functions* that might better capture the variety of human moral judgements. Finally, we conclude and address future possible challenges.

1 Introduction

AI value alignment, the attempt to implement systems adhering to human ethical values has been recognized as highly relevant subtask in AI Safety at an international level and studied by multiple AI and AI Safety researchers across diverse research subareas [Hadfield-Menell *et al.*, 2016; Soares and Fallenstein, 2017; Yudkowsky, 2016] (a review is provided in [Taylor *et al.*, 2016]). Moreover, the need to investigate value alignment has been included in the Asilomar AI Principles [2018] with a worldwide support of researchers from the field. While value alignment has often been tackled using reinforcement learning [Abel *et al.*, 2016] (and also reward modeling [Leike *et al.*, 2018]) or inverse reinforcement learning [Abbeel and Ng, 2004] methods, we focus on the approach to explicitly formulate cardinal ethical utility

functions crafted by (a representation of) society and assisted by science and technology which has been termed *ethical goal functions* [Aliman and Kester, 2019b; Werkhoven *et al.*, 2018]. In order to be able to formulate utility functions that do not violate the ethical intuitions of most entities in a society, these ethical goal functions will have to be a model of human ethical intuitions. This simple but important insight can be derived from the good regulator theorem in cybernetics [Conant and Ross Ashby, 1970] stating that “*every good regulator of a system must be a model of that system*”. We believe that instead of learning models of human intuitions in their apparent complexity and ambiguity, AI Safety research could also make use of the already available scientific knowledge on the nature of human moral judgements and ethical conceptions as made available e.g. by neuroscience and psychology. The human brain did not evolve to facilitate rational decision-making or the experience of emotions, but instead to fulfill the core task of allostasis (anticipating the needs of the body in an environment before they arise in order to ensure growth, survival and reproduction) [Barrett, 2017a; Kleckner *et al.*, 2017]. Thereby, psychological functions such as cognition, emotion or moral judgements are closely linked to the predictive regulation of physiological needs of the body [Kleckner *et al.*, 2017] making it indispensable to consider the embodied nature of morality when aspiring to model it for AI value alignment.

For the purpose of facilitating the injection of requisite knowledge reflecting the variety of human morality in ethical goal functions, Section 2 provides information on the following variety-relevant aspects: 1) the essential role of affect and emotion in moral judgements from a modern constructionist neuroscience and cognitive science perspective followed by 2) dyadic morality as a recent psychological theory on the nature of cognitive templates for moral judgements. In Section 3, we propose first guidelines on how to approximately formulate ethical goal functions using a recently proposed non-normative socio-technological ethical framework grounded in science called *augmented utilitarianism* [Aliman and Kester, 2019a] that might be useful to better incorporate the requisite variety of human ethical intuitions (especially in comparison to classical utilitarianism). Thereafter, we propose how to possibly validate these functions within a socio-technological feedback-loop [Aliman and Kester, 2019b]. Finally, in Section 4, we conclude and specify open challenges

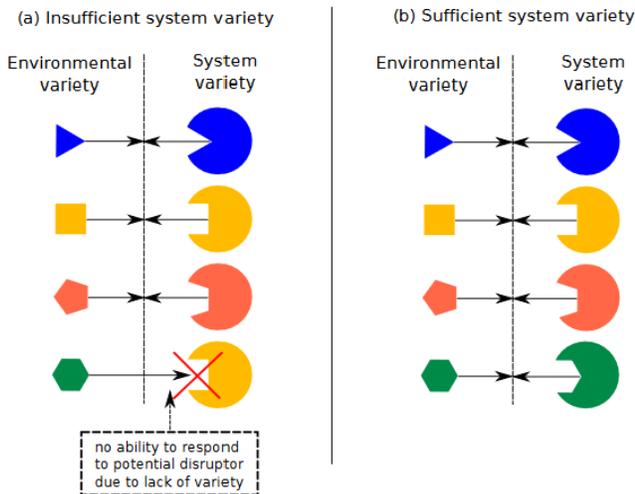


Figure 1: Intuitive illustration for the Law of Requisite Variety. Taken from [Norman and Bar-Yam, 2018].

providing incentives for future work.

2 Variety in Embodied Morality

While value alignment is often seen as a safety problem, it is possible to interpret and reformulate it as a related security problem which might offer a helpful different perspective on the subject emphasizing the need to capture the variety of embodied morality. One possible way to look at AI value alignment is to consider it as being an attempt to achieve advanced AI systems exhibiting adversarial robustness against malicious adversaries attempting to lead the system to action(s) or output(s) that are perceived as violating human ethical intuitions. From an abstract point of view, one could distinguish different means by which an adversary might achieve successful attacks: e.g. 1) by fooling the AI at the perception-level (in analogy to classical adversarial examples [Goodfellow, 2018], this variant has been denoted *ethical adversarial examples* [Aliman and Kester, 2019a]) which could lead to an unethical behavior even if the utility function would have been aligned with human ethical intuitions or 2) simply by disclosing dangerous (certainly unintended from the designer) unethical implications encoded in its utility function by targeting specific mappings from perception to output or action (this could be understood as ethical adversarial examples on the utility function itself). While the existence of point 1) yields one more argument for the importance of research on adversarial robustness at the perception-level for AI Safety reasons [Goodfellow, 2019] and a sophisticated combination of 1) and 2) might be thinkable, our exemplification focuses on adversarial attacks of the type 2).

One could consider the explicitly formulated utility function U as representing a separate model¹ that given a sample, outputs a value determining the perceived ethical desirabil-

¹a conceptually similar separation of objective function model and optimizing agent has been recently performed for reward modeling [Leike *et al.*, 2018]

ity of that sample which should ideally be in line with the society that crafted this utility function. The attacker which has at his disposal the knowledge on human ethical intuitions, can attempt targeted misclassifications at the level of a single sample or at the level of an ordering of multiple samples whereby the ground-truth are the ethical intuitions of most people in a society. The Law of Requisite Variety from cybernetics [Ashby, 1961] states that “*only variety can destroy variety*”, with other words in order to cope with a certain variety of problems or environmental variety, a system needs to exhibit a suitable and sufficient variety of responses. Figure 1 offers an intuitive explanation of this law. Transferring it to the mentioned utility function U , it is for instance conceivable that if U does not encode affective information that might lead to a difference in ethical evaluations, an attacker can easily craft a sample which U might misclassify as ethical or unethical or cause U to generate a total ordering of samples that might appear unethical from the perspective of most people. Given that U does not have an influence on the variety of human morality, the only way to respond to the disturbances of the attacker and reduce the variety of possible undesirable outcomes, is by increasing the own variety – which can be achieved by encoding more relevant knowledge.

2.1 Role of Emotion and Affect in Morality

One fundamental and persistent misconception about human biology (which does not only affect the understanding of the nature of moral judgements) is the assumption that the brain incorporates a layered architecture in which a battle between emotion and cognition is given through the very anatomy of the “*triune brain*” [MacLean, 1990] exhibiting three hierarchical layers: a reptilian brain on top of which an emotional animalistic paleomammalian limbic system is located and a final rational neomammalian cognition layer implemented in the neocortex. This flawed view is not in accordance with neuroscientific evidence and understanding [Barrett, 2017a; Miller and Clark, 2018]. In fact, the assumed reactive and animalistic limbic regions in the brain are predictive (e.g. they send top-down predictions to more granular cortical regions), control the body as well as attention mechanisms while being the source of the brain’s internal model of the body [Barrett and Simmons, 2015; Barrett, 2017b].

Emotion and cognition do not represent a dichotomy leading to a conflict in moral judgements [Helion and Pizarro, 2015]. Instead, the distinction between the experience of an instance of a concept as belonging to the category of emotions versus the category of cognition is grounded in the focus of attention of the brain [Barrett *et al.*, 2015] whereby “*the experience of cognition occurs when the brain foregrounds mental contents and processes*” and “*the experience of emotion occurs when, in relation to the current situation, the brain foregrounds bodily changes*” [Hoemann and Barrett, 2019]. The mental phenomenon of actively dynamically simulating different alternative scenarios (including anticipatory emotions) has also been termed conceptual consumption [Gilbert and Wilson, 2007] and plays a role in decision-making and moral reasoning. While emotions are discrete constructions of the human brain, core affect allows a low-dimensional experience of interoceptive sensations (sensory array from

within the body) and is a continuous property of consciousness with the dimensions of valence (pleasantness/unpleasantness) and arousal (activation/deactivation) [Kleckner *et al.*, 2017]. It has been argued that core affect provides a basis for moral judgements in which different events are qualitatively compared to each other [Cabanac, 2002]. Like other constructed mental states, moral judgements involve domain-general brain processes which simply put combine 1) the interoceptive sensory array, 2) the exteroceptive sensory inputs from the environment and 3) past experience/ knowledge for a goal-oriented situated conceptualization (as tool for allostasis) [Oosterwijk *et al.*, 2012]. From these key constituents of mental constructions one can extract the following: concepts (including morality) are *perceiver-dependent* and *time-dependent*. Thereby, affect, (but not emotion [Cameron *et al.*, 2015]) is a necessary ingredient of every moral judgement. More fundamentally, “*the human brain is anatomically structured so that no decision or action can be free of interoception and affect*” [Barrett, 2017a] – this includes any type of thoughts that seem to correspond to the folk terms of “rational” and “cold”. Therefore, a utility function without affect-related parameters might not exhibit a sufficient variety and might lead to the violation of human ethical intuitions.

Morality cannot be separated from a model of the body, since the brain constructs the human perception of reality based on what seems of importance to the brain for the purpose of allostasis which is inherently strongly linked to interoception [Barrett, 2017a]. Interestingly, even the imagination of future not yet experienced events is facilitated through situated recombinations of sensory-motor and affective nature in a similar way as the simulation of actually experienced events [Addis, 2018]. To sum up, there is no battle between emotion and cognition in moral judgements. Moreover, there is also no specific moral faculty in the brain, since moral judgements are based on domain-general processes within which affect is always involved to a certain degree. One could obtain insufficient variety in dealing with an adversary crafting ethical adversarial examples on a utility model U if one ignores affective parameters. Further crucial parameters for ethical utility functions could be e.g. of cultural, social and socio-geographical nature.

2.2 Variety through “Dyadicness”

The psychological theory of dyadic morality [Schein and Gray, 2018] posits that moral judgements are based on a fuzzy cognitive template and related to the perception of an intentional agent (iA) causing damage (d) to a vulnerable patient (vP) denoted $iA \xrightarrow{d} vP$. More precisely, the theory postulates that the perceived immorality of an act is related to the following three elements: norm violations, negative affect and importantly perceived harm. According to a study, the reaction times in describing an act as immoral predict the reaction times in categorizing the same act as harmful [Schein and Gray, 2015]. The combination of these basic constituents is suggested to lead to the emergence of a rich diversity of moral judgements [Gray *et al.*, 2017]. *Dyadicness* is understood as a continuum predicting the condemnation of moral acts. The more a human entity perceives an intentional agent inflicting damage to a vulnerable patient, the more immoral

this human perceives the act. As stated by Schein and Gray, the dyadic harm-based cognitive template “*is rooted in innate and evolved processes of the human mind; it is also shaped by cultural learning, therefore allowing cultural pluralism*”. Importantly, the nature of this cognitive template reveals that moral judgements besides being perceiver-dependent, might vary across diverse parameters such as especially e.g. in relation to the perception of agent, act and patient in the outcome of the action. Further, the theory also foresees a possible time-dependency of moral judgements by introducing the concept of a *dyadic loop*, a feedback cycle resulting in an iterative polarization of moral judgements through social discussion modulating the perception of harm as time goes by. Overall, moral judgements are understood as constructions in the same way visual perception, cognition or emotion are constructed by the human mind. Similarly to the existence of variability in visual perception, variability in morality is the norm which often leads to moral conflicts [Schein *et al.*, 2016]. However, the understanding that humans share the same harm-based cognitive template for morality has been described as reflecting “*cognitive unity in the variety of perceived harm*” [Schein and Gray, 2018].

Analyzing the cognitive template of dyadic morality, one can deduce that human moral judgements do not only consider the outcome of an action as prioritized by consequentialist frameworks like classical utilitarianism, nor do they only consider the state of the agent which is in the focus of virtue ethics. Furthermore, as opposed to deontological ethics, the focus is not only on the nature of the performed action. The main implications for the design of utility functions that should ideally be aligned with human ethical values, is that they might need to encode information on agent, action, patient as well as on the perceivers – especially with regard to the cultural background. This observation is fundamental as it indicates that one might have to depart from classical utilitarian utility functions $U(s')$ which are formulated as total orders at the abstraction level of outcomes i.e. states (of affairs) s' . In line with this insight, is the context-sensitive and perceiver-dependent type of utility functions considering agent, action and outcome which has been recently proposed within a novel ethical framework denoted *augmented utilitarianism* [Aliman and Kester, 2019a] (abbreviated with AU in the following). Reconsidering the dyadic morality template $iA \xrightarrow{d} vP$, it seems that in order to better capture the variety of human morality, utility functions – now transferring it to the perspective of AI systems – would need to be at least formulated at the abstraction level of a *perceiver-dependent* evaluation of a transition $s \xrightarrow{a} s'$ leading from a state s to a state s' via an action a . We encode the required novel type of utility function with $U_x(s, a, s')$ with x denoting a specific perceiver. This formulation could enable an AI system implemented as utility maximizer to jointly consider parameters specified by a perceiver which are related to its perception of agent, the action and the consequences of this action on a patient. Since the need to consider time-dependency has been formulated, one would consequently also require to add the time dimension to the arguments of the utility function leading to $U_x((s, a, s'), t)$.

3 Approximating Ethical Goal Functions

While the psychological theory of dyadic morality was useful to estimate the abstraction level at which one would at least have to specify utility functions, the closer analysis on the nature of the construction of mental states performed in Section 2, abstractly provides a superset of primitive relevant parameters that might be critical elements of every moral judgement (being a mental state). Given a perceiver x , the components of this set are the following subsets: 1) parameters encoding the interoceptive sensory array B_x (from within the body) which are accessible to the human consciousness via the low-dimensional core affect, 2) the exteroceptive sensory array E_x encoding information from the environment and 3) the prior experience P_x encoding memories. Moreover, these set of parameters obviously vary in time. However, to simplify, it has been suggested within the mentioned AU framework, that ethical goal functions will have to be updated regularly (leading to a so-called socio-technological feedback-loop [Aliman and Kester, 2019b]) in the same way as votes take place at regular intervals in a democracy. One could similarly assume that this regular update will be sufficient to reflect a relevant change in moral opinion and perception.

3.1 Injecting Requisite Variety in Utility

For simplicity, we assume that the set of parameters B_x , P_x and E_x are invariant during the utility assignment process in which a perceiver x has to specify the ethical desirability of a transition $s \xrightarrow{a} s'$ by mapping it to a cardinal value $U_x(s, a, s')$ obtained by applying a not-nearer defined type of scientifically determined transformation v_x (chosen by x) on the mental state of x . This results in the following naive and simplified mapping however adequately reflecting the property of *mental-state-dependency* formulated in the AU framework (the required dependency of ethical utility functions on parameters of the own mental state function m_x in order to avoid perverse instantiation scenarios [Aliman and Kester, 2019a]):

$$U_x(s, a, s') = v_x(m_x((s, a, s'), B_x, P_x, E_x)) \quad (1)$$

Conversely, the utility function of classical utilitarianism is only defined at the impersonal and context-independent abstraction level of $U(s')$ which has been argued to lead to both *perverse* instantiation problem but also to the *repugnant* conclusion and related impossibility theorems in population ethics for consequentialist frameworks which do not apply to mental-state-dependent utility functions [Aliman and Kester, 2019a]. The idea to restrict human ethical utility functions to the considerations of outcomes of actions alone – ignoring affective parameters of the own current self – as practiced in classical utilitarianism while later referring to the resulting total orders with emotionally connoted adjectives such as “repugnant” or “perverse” has been termed the *perspectival fallacy of utility assignment* [Aliman and Kester, 2019b]. The use of consequentialist utility functions affected by the impossibility theorems of Arrhenius [2000] has been justifiably identified by Eckersley [2018] as a safety risk if used in AI systems without more ado. It seems that the isolated consideration of outcomes of actions (for consequentialism

or actions (for deontological ethics) or the involved agents (for virtue ethics) does not represent a good model of human ethical intuitions. It is conceivable, that if a utility model U is defined as utility function $U(s')$, the model cannot possibly exhibit a sufficient variety and might more likely violate human ethical intuitions than if it would be implemented as a context-sensitive utility function $U_x(s, a, s')$. (Beyond that, it has been argued that consequentialism implies the rejection of “*dispositions and emotions, such as regret, disappointment, guilt and resentment*” from “rational” deliberation [Verbeek, 2001] and should i.a. for this reason be disentangled from the notion of rationality for which it cannot represent a plausible requirement.)

It is noteworthy that in the context of reinforcement learning (e.g. in robotics) different types of reward functions are usually formulated ranging from $R(s')$ to $R(s, a, s')$. For the purpose of ethical utility functions for advanced AI systems in critical application fields, we postulate that one does not have the choice to specify the abstraction level of the utility function, since for instance $U(s')$ might lead to safety risks. Christiano et al. [2017] considered the elicitation of human preferences on trajectory (state-action pairs) segments of a reinforcement learning agent i.a. realized by human feedback on short movies. For the purpose of utility elicitation in an AU framework exemplarily using a naive model as specified in equation (1), people will similarly have to assign utility to a movie representing a transition in the future (either in a mental mode or augmented by technology such as VR or AR [Aliman and Kester, 2019b]). However, it is obvious that this naive utility assignment would not scale in practice. Moreover, it has not yet been specified how to aggregate ethical goal functions at a societal level. In the following Subsection 3.2, we will address these issues by proposing a practicable approximation of the utility function in (1) and a possible societal aggregation of this approximate solution.

3.2 Approximation, Aggregation and Validation

So far, it has been stated throughout the paper that one has to adequately increase the variety of a utility function meant to be ethical in order to avoid violations of human ethical intuitions and vulnerability to attackers crafting ethical adversarial examples against the model. However, it is important to note that despite the negatively formulated motivation of the approach, the aim is to craft a utility model U which represents a better model of human ethical intuitions in general, thus ranging from samples that are perceived as highly unethical to those that are assigned a high ethical desirability. In order to craft practical solutions that lead to optimal results, it might be advantageous to perform a thought experiment imagining a utopia and from that impose practical constraints on its viability. It might not seem realistic to liberate a future *utopia 1* as a sustainable society which is stable across a very large time interval in which every human being acts according to the ethical intuitions of all humans including the own and every artificial intelligent system fulfills the ethical intuitions of all humans. However, it seems more likely that within a *utopia 2* being a stable society in which every human achieves a high level of a scientific definition of well-being (such as e.g. PERMA [Seligman, 2012])

with artificial agents acting as to maximize context-sensitive utility according to which (human or artificial) agents promoting the (measurable) well-being of human patients is regarded as the most utile type of events, the ethical intuitions of humans might tend to get closer to each other. The reason being that the variety of human moral judgements might interestingly *decrease* since it is conceivable that they will tend to exhibit more similar prior experiences (all imprinted by well-being) and have more similar environments (full of stable people with a high level of well-being). The main factor drawing differences could be the body – especially biological factors. However, the parameters related to interoception might be closer to each other, since all humans exhibit a high level of well-being which classically includes frequent positive affect. It is conceivable that with time, such a society could converge towards the utopia 1.

In the following, we will denote the mentioned utopia-related ideal cognitive template of a (human or artificial) agent A performing an act w that contributes to the well-being of a human patient P with $A \xrightarrow{w} P$ in analogy to the cognitive template of dyadic morality. (Thereby, $A \xrightarrow{w} P$ is perceiver-dependent i.a. because psychological measures of well-being include subjective and self-reported elements such as e.g. life satisfaction or furthermore positive emotions [Seligman, 2012].) Augmented utilitarianism foresees the need to at least depict a final goal at the abstraction level of a perceiver-dependent function on a transition as reflected in $U_x(s, a, s')$. The ideal cognitive template $A \xrightarrow{w} P$ formulated for utopia 2 by which it has been argued that a decrease in the variety of human morality might be achievable in the long-term exhibits an abstraction level that is compatible with $U_x(s, a, s')$.

A thinkable strategy for the design of a utility model U that is robust against ethical adversarial examples and a model of human ethical intuitions is to try to adequately increase its variety using relevant scientific knowledge and to complementarily attempt to decrease the variety of human moral judgements for instance by considering $A \xrightarrow{w} P$ as high-level final goal such that the described utopia 2 ideally becomes a self-fulfilling prophecy. For it to be realizable in practice, we suggest that the appropriateness of a given aggregated societal ethical goal function could be approximately validated against its quantifiable impact on well-being for society across the time dimension. Since it seems however unfeasible to directly map all important transitions of a domain to their effect on the well-being of human entities, we propose to consider perceiver-specific and domain-specific utility functions indicating combined preferences that each perceiver x considers to be relevant for well-being from the viewpoint of x himself in that specific domain. For these combined utility functions to be grounded in science, they will have to be based on scientifically measurable parameters. We postulate that a possible aggregation at a societal level could be performed by the following steps: 1) agreement on a common validation measure of an ethical goal function (for instance the temporal development of societal satisfaction with AI systems in a certain domain or with future AGI systems, their aptitude to contribute to sustainable well-

being), 2) agreement on *superset* O of scientifically measurable and relevant parameters (encoding e.g. affective, dyadic, cultural, social, political, socio-geographical but importantly also law-relevant information) that are considered as important across the whole society, 3) specification of personal utility functions for each member n of a society of N members allowing personalized and tailored combinations of a subset of O , 4) aggregation to a societal ethical goal function $U_{Total}(s, a, s')$. Taken together, these considerations lead us to the following possible approximation for an aggregated societal ethical goal function given a domain:

$$U_{Total}(s, a, s') = \sum_{n=1}^N \sum_{i=1}^j w_i^n f_{fi}^n(C_i) \quad (2)$$

with N standing for the number of participating entities in society, $C_i = (p_{i1}, p_{i2}, \dots, p_{im})$ being a cluster of $m \geq 1$ correlated parameters (whereby independent factors are assigned an own cluster each) and f representing a set of preference functions (*form functions*). For instance $f = \{f_1, f_2, \dots, f_f\}$ where f_1 could be a linear transformation, f_2 a concave, f_3 a convex preference function and so on. Each entity n assigns a weight w_i^n to a form function f_{fi}^n applied to a cluster of parameters C_i whereby $\sum_{i=1}^j w_i^n = 1$. We define $O = \{C_1, C_2, \dots\}$ as the superset of all parameters considered in the overall aggregated utility function. Moreover, $a \in A$ with A representing the foreseen discrete action space at the disposal of the AI. (It is important to note that while the AI could directly perform actions in the environment, it could also be used for policy-making and provide plans for human agents.) Further, we consider a continuous state space with the states s and $s' \in S = \mathbb{R}^{|O|}$. Other aspects including e.g. legal rules and norms on the action space can be imposed as constraints on the utility function. In a nutshell, the utility aggregation process can be understood as a voting process in which each participating individual n distributes his vote across scientifically measurable clusters of parameters C_i on which he applies a preference function f_{fi}^n to which weights w_i^n are assigned as identified as relevant by n given a to be approximated high-level societal goal (such as $A \xrightarrow{w} P$). In short, people do not have to agree on personal preferences and weightings, but only on a superset of acceptable parameters, an aggregation method and an overall validation measure. (Note that instead of involving society as a whole for each domain, the utility elicitation procedure can as well be approximated by a transdisciplinary set of representative experts (e.g. from the legislative) crafting *expert ethical goal functions* that attempt to ideally emulate $U_{Total}(s, a, s')$).

Finally, it is important to note that the societal ethical goal function specified in (2) will need to be updated (and evaluated) at regular intervals due to the mental-state-dependency of utility entailing time-dependency [Aliman and Kester, 2019a]. This leads to the necessity of a socio-technological feedback-loop which might concurrently offer the possibility of a *dynamical ethical enhancement* [Aliman and Kester, 2019b; Werkhoven *et al.*, 2018]. Pre-deployment, one could in the future attempt a validation via selected preemptive simulations [Aliman and Kester, 2019b] in which (a represen-

tation of) society experiences simulations of future events (s, a, s') as movies, immersive audio-stories or later in VR and AR environments. During these experiences, one could approximately measure the temporal profile of the so-called *artificially simulated future instant utility* [Aliman and Kester, 2019b] denoted $U_{TotalAS}$ being a potential constituent of future well-being. Thereby, $U_{TotalAS}$ refers to the instant utility [Kahneman *et al.*, 1997] experienced during a technology-aided simulation of a future event whereby instant utility refers to the affective dimension of valence at a certain time t . The temporal integral that a measure of $U_{TotalAS}$ could approximate is specified as:

$$U_{TotalAS}(s, a, s') \approx \sum_{n=1}^N \int_{t_0}^T I_n(t) dt \quad (3)$$

with t_0 referring to the starting point of experiencing the simulation of the event (s, a, s') augmented by technology (movie, audio-story, AR, VR) and T the end of this experience. $I_n(t)$ represents the valence dimension of core affect experienced by n at time t . Finally, post-deployment, the ethical goal function of an AI system can be validated using the validation measure agreed upon before utility aggregation (such as the temporal development of societal-level satisfaction with an AI system, well-being or even the perception of dyadicness) that has to be a priori determined.

4 Conclusion and Future Work

In this paper, we motivated the need in AI value alignment to attempt to model utility functions capturing the variety of human moral judgements through the integration of relevant scientific knowledge – especially from neuroscience and psychology – (instead of learning) in order to avoid violations of human ethical intuitions. We reformulated value alignment as a security task and introduced the requirement to increase the variety within classical utility functions positing that a utility function which does not integrate affective and perceiver-dependent dyadic information does not exhibit sufficient variety and might not exhibit robustness against corresponding adversaries. Using augmented utilitarianism as a suitable non-normative ethical framework, we proposed a methodology to implement and possibly validate societal perceiver-dependent ethical goal functions with the goal to better incorporate the requisite variety for AI value alignment.

In future work, one could extend and refine the discussed methodology, study a more systematic validation approach for ethical goal functions and perform first experimental studies. Moreover, the “*security of the utility function itself is essential, due to the possibility of its modification by malevolent actors during the deployment phase*” [Aliman and Kester, 2019a]. For this purpose, a blockchain-based solution might be advantageous. In addition, it is important to note that even with utility functions exhibiting a sufficient variety for AI value alignment, it might still be possible for a malicious attacker to craft adversarial examples against a utility maximizer at the perception-level which might lead to unethical behavior. Besides that, one might first need to perform policy-by-simulation [Werkhoven *et al.*, 2018] prior to the deployment of advanced AI systems equipped with ethical goal

functions for safety reasons. Last but not least, the usage of ethical goal functions might represent an interesting approach to the AI coordination subtask in AI Safety, since an international use of this method might contribute to reduce the AI race to the problem-solving ability dimension [Aliman and Kester, 2019b].

References

- [Abbeel and Ng, 2004] Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 1. ACM, 2004.
- [Abel *et al.*, 2016] David Abel, James MacGlashan, and Michael L Littman. Reinforcement learning as a framework for ethical decision making. In *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*, 2016.
- [Addis, 2018] Donna Rose Addis. Are episodic memories special? On the sameness of remembered and imagined event simulation. *Journal of the Royal Society of New Zealand*, 48(2-3):64–88, 2018.
- [Aliman and Kester, 2019a] Nadisha-Marie Aliman and Leon Kester. Augmented Utilitarianism for AGI Safety. In *International Conference on Artificial General Intelligence*, page to appear. Springer, 2019.
- [Aliman and Kester, 2019b] Nadisha-Marie Aliman and Leon Kester. Transformative AI Governance and AI-Empowered Ethical Enhancement Through Preemptive Simulations. *Delphi - Interdisciplinary Review of Emerging Technologies*, 2(1):23–29, 2019.
- [Arrhenius, 2000] Gustaf Arrhenius. An impossibility theorem for welfarist axiologies. *Economics & Philosophy*, 16(2):247–266, 2000.
- [Ashby, 1961] W Ross Ashby. *An introduction to cybernetics*. Chapman & Hall Ltd, 1961.
- [Asilomar, 2018] AI Asilomar. Principles.(2017). In *Principles developed in conjunction with the 2017 Asilomar conference [Benevolent AI 2017]*, 2018.
- [Barrett and Simmons, 2015] Lisa Feldman Barrett and W Kyle Simmons. Interoceptive predictions in the brain. *Nature Reviews Neuroscience*, 16(7):419, 2015.
- [Barrett *et al.*, 2015] Lisa Feldman Barrett, Christine D Wilson-Mendenhall, and Lawrence W Barsalou. The conceptual act theory: A roadmap. pages 83–110, 2015.
- [Barrett, 2017a] Lisa Feldman Barrett. *How emotions are made: The secret life of the brain*. Houghton Mifflin Harcourt, 2017.
- [Barrett, 2017b] Lisa Feldman Barrett. The theory of constructed emotion: an active inference account of interoception and categorization. *Social cognitive and affective neuroscience*, 12(1):1–23, 2017.
- [Cabanac, 2002] Michel Cabanac. What is emotion? *Behavioural processes*, 60(2):69–83, 2002.
- [Cameron *et al.*, 2015] C Daryl Cameron, Kristen A Lindquist, and Kurt Gray. A constructionist review of

- morality and emotions: No evidence for specific links between moral content and discrete emotions. *Personality and Social Psychology Review*, 19(4):371–394, 2015.
- [Christiano *et al.*, 2017] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, pages 4299–4307, 2017.
- [Conant and Ross Ashby, 1970] Roger C Conant and W Ross Ashby. Every good regulator of a system must be a model of that system. *International journal of systems science*, 1(2):89–97, 1970.
- [Eckersley, 2018] Peter Eckersley. Impossibility and Uncertainty Theorems in AI Value Alignment (or why your AGI should not have a utility function). *arXiv preprint arXiv:1901.00064*, 2018.
- [Gilbert and Wilson, 2007] Daniel T Gilbert and Timothy D Wilson. Propection: Experiencing the future. *Science*, 317(5843):1351–1354, 2007.
- [Goodfellow, 2018] Ian Goodfellow. Defense Against the Dark Arts: An overview of adversarial example security research and future research directions. *arXiv preprint arXiv:1806.04169*, 2018.
- [Goodfellow, 2019] Ian Goodfellow. Adversarial Robustness for AI Safety. <https://safeai.webs.upv.es/wp-content/uploads/2019/02/2019-01-27-goodfellow.pdf>, 2019.
- [Gray *et al.*, 2017] Kurt Gray, Chelsea Schein, and C Daryl Cameron. How to think about emotion and morality: circles, not arrows. *Current opinion in psychology*, 17:41–46, 2017.
- [Hadfield-Menell *et al.*, 2016] Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. Cooperative inverse reinforcement learning. In *Advances in neural information processing systems*, pages 3909–3917, 2016.
- [Helion and Pizarro, 2015] Chelsea Helion and David A Pizarro. Beyond dual-processes: the interplay of reason and emotion in moral judgment. *Handbook of neuroethics*, pages 109–125, 2015.
- [Hoemann and Barrett, 2019] Katie Hoemann and Lisa Feldman Barrett. Concepts dissolve artificial boundaries in the study of emotion and cognition, uniting body, brain, and mind. *Cognition and Emotion*, 33(1):67–76, 2019. PMID: 30336722.
- [Kahneman *et al.*, 1997] Daniel Kahneman, Peter P Wakker, and Rakesh Sarin. Back to Bentham? Explorations of experienced utility. *The quarterly journal of economics*, 112(2):375–406, 1997.
- [Kleckner *et al.*, 2017] Ian R Kleckner, Jiahe Zhang, Alexandra Touroutoglou, Lorena Chanes, Chenjie Xia, W Kyle Simmons, Karen S Quigley, Bradford C Dickerson, and Lisa Feldman Barrett. Evidence for a large-scale brain system supporting allostasis and interoception in humans. *Nature human behaviour*, 1(5):0069, 2017.
- [Leike *et al.*, 2018] Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*, 2018.
- [MacLean, 1990] Paul D MacLean. *The triune brain in evolution: Role in paleocerebral functions*. Springer Science & Business Media, 1990.
- [Miller and Clark, 2018] Mark Miller and Andy Clark. Happily entangled: prediction, emotion, and the embodied mind. *Synthese*, 195(6):2559–2575, 2018.
- [Norman and Bar-Yam, 2018] Joseph Norman and Yaneer Bar-Yam. Special Operations Forces: A Global Immune System? In *International Conference on Complex Systems*, pages 486–498. Springer, 2018.
- [Oosterwijk *et al.*, 2012] Suzanne Oosterwijk, Kristen A Lindquist, Eric Anderson, Rebecca Dautoff, Yoshiya Moriguchi, and Lisa Feldman Barrett. States of mind: Emotions, body feelings, and thoughts share distributed neural networks. *NeuroImage*, 62(3):2110–2128, 2012.
- [Schein and Gray, 2015] Chelsea Schein and Kurt Gray. The unifying moral dyad: Liberals and conservatives share the same harm-based moral template. *Personality and Social Psychology Bulletin*, 41(8):1147–1163, 2015.
- [Schein and Gray, 2018] Chelsea Schein and Kurt Gray. The theory of dyadic morality: Reinventing moral judgment by redefining harm. *Personality and Social Psychology Review*, 22(1):32–70, 2018.
- [Schein *et al.*, 2016] Chelsea Schein, Neil Hester, and Kurt Gray. The visual guide to morality: Vision as an integrative analogy for moral experience, variability and mechanism. *Social and Personality Psychology Compass*, 10(4):231–251, 2016.
- [Seligman, 2012] Martin EP Seligman. *Flourish: A visionary new understanding of happiness and well-being*. Simon and Schuster, 2012.
- [Soares and Fallenstein, 2017] Nate Soares and Benya Fallenstein. Agent foundations for aligning machine intelligence with human interests: a technical research agenda. In *The Technological Singularity*, pages 103–125. Springer, 2017.
- [Taylor *et al.*, 2016] Jessica Taylor, Eliezer Yudkowsky, Patrick LaVictoire, and Andrew Critch. Alignment for advanced machine learning systems. *Machine Intelligence Research Institute*, 2016.
- [Verbeek, 2001] Bruno Verbeek. Consequentialism, rationality and the relevant description of outcomes. *Economics & Philosophy*, 17(2):181–205, 2001.
- [Werkhoven *et al.*, 2018] Peter Werkhoven, Leon Kester, and Mark Neerinx. Telling autonomous systems what to do. In *Proceedings of the 36th European Conference on Cognitive Ergonomics*, page 2. ACM, 2018.
- [Yudkowsky, 2016] Eliezer Yudkowsky. The AI Alignment Problem: Why it is Hard, and Where to Start. *Symbolic Systems Distinguished Speaker*, 2016.