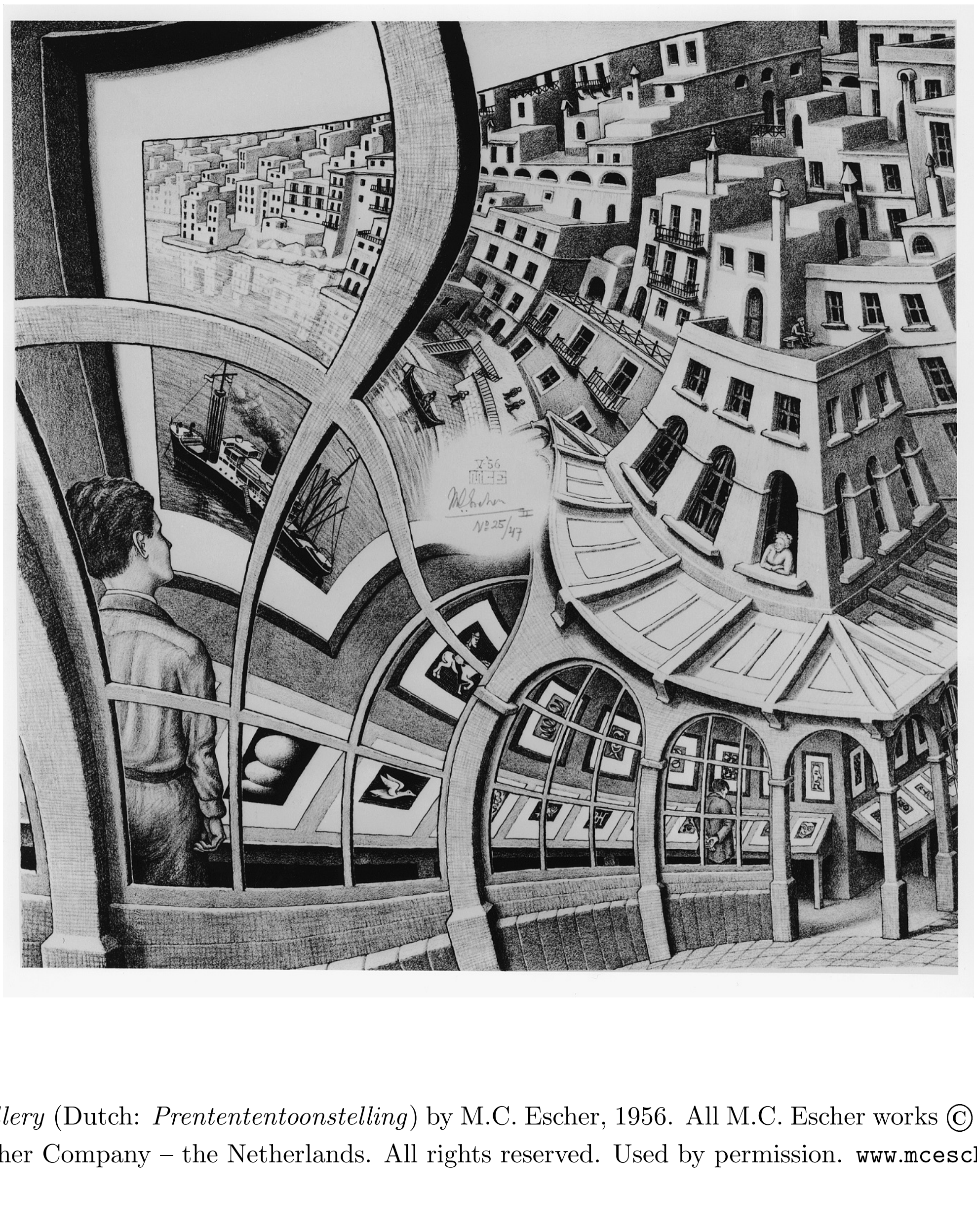
Ph.D. Thesis

Hybrid Cognitive-Affective Strategies for AI Safety

Nadisha-Marie Aliman



Print Gallery (Dutch: *Prententententoonstelling*) by M.C. Escher, 1956. All M.C. Escher works © 2020 The M.C. Escher Company – the Netherlands. All rights reserved. Used by permission. www.mcescher.com

Hybrid Cognitive-Affective Strategies for AI Safety

Hybride cognitief-affectieve strategieën voor AI-veiligheid
(met een samenvatting in het Nederlands)

PROEFSCHRIFT
ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof. dr. H.R.B.M. Kummeling, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen
op
2 december 2020 te 12.45 uur

door

Nadisha-Marie Aliman
geboren op 23 mei 1992
te Bochum, Duitsland

Promotoren: Prof. dr. P.J. Werkhoven
Prof. dr. ir. J. Masthoff

Contents

Chapter 1

Introduction

While it is conceivable that artificial intelligence (AI) could have a tremendously beneficial future impact on society, the associated ethical and safety-relevant ramifications need to be addressed given their potential range. In this light, the field of *AI safety* has been recognized as a critical academic research direction at an international level over the intervening years [33, 223, 392, 400, 456]. Etymologically speaking, "AI safety" represents a shortened form for the longer phrase "artificial intelligence safety and security engineering". Both terms were introduced in 2010 in a peer-reviewed form by Roman Yampolskiy [451], one of the early advocates of the underlying relatively young field. Prior to that, terms like "machine ethics" [308] and "friendly AI" [453] were utilized to refer to different related subtopics. However, this thesis is rather premised on AI safety in its property of being framed from a cybersecurity-oriented and risk-centered perspective [338, 446, 451]. Although there is currently no consensus on a formal definition for AI safety, its initial research objectives can be outlined when considering the main task it addressed. Hitherto, AI safety research was predominantly concentrated on the following *twofold* key issue: solving both the *value alignment problem* [456] and the *control problem* [83]. More precisely, the former refers to the problem on *how to implement AIs aligned with human ethical values* and the latter to the connected issue on *how to build AI systems that will not harm humans.* While a more comprehensive and risk-centered account going beyond this twofold view on AI safety is specified in a few paragraphs, the main research question of this thesis can already be indicated as follows: *"what types of scientifically grounded strategies can be employed to facilitate AI safety?"*.

Given the steadily progressing capabilities of AI, its integration into decision-making processes and its application in areas affording an increasing degree of autonomy, requirements for AI safety need to already be urgently identified at this stage. In order to do justice to the underlying broad and complex research issue, this theoretical and analytical thesis devises an integrated approach via a *transdisciplinary* methodology dovetailing considerations from diverse scientific fields. For this purpose, we establish a set of *hybrid* and *cognitive-affective* strategical clusters to foster AI safety – a field which we remodel throughout the thesis and whose problem space we shift and recontextualize. A priori, the necessity to consider "hybrid" solutions emanates from the realization that improvement and enhancement measures for safer AI systems cannot be solely applied to the artificial system in isolation [428]. Instead, AI has to be considered in the wide socio-technological context of its deployment amidst *human* entities [30]. As a result, it is crucial to account for human cognitive dispositions when designing AI safety strategies for the implementation of value-aligned and controllable AIs. Building on this, AI safety strategies need to more specifically encompass "cognitive-affective" aspects due to the inherently affective nature [47, 111, 150, 263] of human cognition. With that in mind, this thesis harnesses a quintessentially transdisciplinary and hybrid cognitive-affective approach to AI safety in order to obviate methodological blind spots. Simultaneously, this thesis facilitates a coalescence of both short-term and long-term AI safety.

Given the breadth of the topic, we paradigmatically take the meaningful control of "intelligent systems"[1] as point of departure – an already practically relevant and graspable contemporary challenge [428] encompassing both control and value alignment issues. After determining AI safety strategies for the meaningful control of intelligent systems from Chapter 2 to 9, we analyze their transferability to other sorts of AI systems in Chapter 10. In the course of this, we identify the fundamental necessity of an ethically-relevant distinction between the yet to be described "Type I" systems and "Type II" systems. On this view, all present-day AIs (including intelligent systems) represent a subset of Type I AI with Type I and Type II AI systems representing disjunct sets. To put it very simply, hypothetical not yet implemented Type II AI systems can be characterized as exhibiting the ability to *consciously* create and understand *explanatory knowledge.* Explanatory knowledge creation can be simply seen as a process that brings forth explanations. In this connection, Deutsch describes an explanation as being a *"statement about what is there, what it does, and how and why"* [136]. Thereby, note that we utilize the term "Type II systems" as a general substrate-independent denotation for systems capable of consciously creating and understanding explanatory knowledge. Obviously, human beings represent Type II systems whereby *"the processes of constructing and understanding explanations are intrinsic to our mental lives from an early age"* [260]. When considering Type II systems, it is important not to overlook the element of *conscious understanding* implied. Thus, by way of example, a present-day recommender AI system or intelligent system generating a set of data that human entities conceive of as explanations, does *not* represent a Type II system. It instead clearly corresponds to a Type I system.

[1] Intelligent systems are often referred to as "autonomous systems". In this thesis, we opt for the former term to emphasize that these Type I AI systems are *not* autonomously and independently setting own intentional goals. While intelligent systems are described to feature the ability to independently perform the OODA-loop (Observe, Orient, Decide, Act), their decision component is fully pre-determined by human-defined goal settings [153]. Hence, "autonomous vehicles" are understood as intelligent systems.

While this thesis predominantly focuses on Type I AI with Type II AI only covered in the last chapters, we briefly comment on Type II AI for more clarity in the delineation. Importantly, what we refer to as Type II AI is also *not* characterized by "superintelligence"[2] as frequently pronounced in the AI safety literature. Instead, we zero in on conscious understanding and explanatory knowledge linked to the creation of explanations. Thereby, we emphasize the *why* aspect in Type II systems and the hereto associated active elucidation and conscious understanding-why of novel previously unknown perspectives on the external and internal milieu of the system at different spatiotemporal scales. For instance, hypothetical future Type II AI could have the capability to construct and understand novel concepts in the domain of ideas regarding science, technology and philosophy but also morality and social reality including the underlying cognitive-affective, embodied and participatory sense-making[3]. In contrast to this, Type I AI is constrained to the specifications of its implementation in the sense that it can neither transcend those consciously nor understand and transform its own nature. By way of example, present-day run-time adaptive intelligent systems can be designed to dynamically adapt and "transform" in line with given specifications such as utility functions [261]. However, those Type I systems could not consciously aim at transforming themselves nor could they develop a conscious understanding of what it means to be a system that transforms itself and consciously engage in disobedience. Yet, all together, even if hypothetical Type II AI might seem futuristic nowadays, it seems responsible to theoretically consider its opportunities, risks and implications since its implementation is physically speaking possible [135, 136] and is not prohibited by any law of nature. Moreover, even if humanity would not succeed to craft a Type II AI in the long-term, the intense examination of this topic might provide scientific insights of inherent value for human (self-)knowledge. Hence, while the focus of this thesis is set on the more tangible Type I AI safety, we also briefly elaborate on Type II AI especially in Chapter 10 and 12.

Initially, AI safety was focused on long-term safety considerations related to a scenario termed "intelligence explosion" [394, 457] or "technological singularity" [451] – a putative moment at which AI abruptly surpasses human intelligence. (Already in 1965, such an event was speculated to imply an ultraintelligent recursively self-improving AI [197].) While it was acknowledged that the exact date of occurrence of this phenomenon is unknown, researchers often referred to a heuristic prediction of Ray Kurzweil according to which machine intelligence would surpass its human counterpart at around 2045 [272]. Against this backdrop, AI safety often framed the value alignment and the control prob-

[2] Mostly, the use of "superintelligence" in AI safety refers to a super-human problem-solving ability that *"exceeds the cognitive performance of humans in virtually all domains of interest"* [83]. In this word usage in the literature, it prevalently corresponds to a Type I AI and does not even imply artificial consciousness, let alone conscious explanatory knowledge creation.

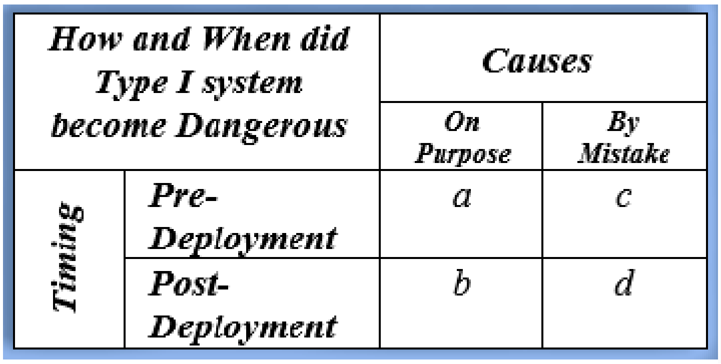
[3] Likewise, cognition has been depicted as *"an embodied, enactive, affective process involving cultural affordances"* [420] and has been linked to social participatory sense-making [130, 299].

lem in this specific context of conjectured superintelligent AI systems. Thereby, this type of AI system was understood to represent a potential enabler of existential risks for humanity [28] given that a *"single failure of a superintelligent system may cause a catastrophic event without a chance for recovery"* [451]. Proposed solutions to solve the control problem for superintelligent systems ranged from building an artificial superintelligence constrained to be a disembodied question-answering system denoted "oracle AI" [28] over initiating a "controlled intelligence explosion" [311] to confining the system within an "AI box" [411, 442]. Furthermore, in order to solve both control and value alignment problem, it was suggested to a priori restrict any goal-directed behavior in superintelligent systems to durable human-crafted and human-friendly goals [441]. Proponents of related approaches often discarded any AI research aimed at artificial systems with a set of *own* motivations and intentions arguing away any philosophical debates related to associated moral rights [443]. Beyond that, other researchers assumed the unachievability of mathematico-logical guarantees for the safety of human-level and superintelligent AI systems and proposed to focus on early educational measures [71]. In addition, mathematical work focused on artificial general intelligence (AGI) formulated within the reinforcement learning paradigm proposed a set of *"partial solutions"* [159] to AI safety.

Whereas the mentioned early approaches to AI safety focused on highly advanced AI systems exhibiting human-level intelligence (and beyond), a short-term oriented branch of this research field emerged which brings into focus present-day AI systems. Thereby, the emphasis is frequently set on unintentionally occurring safety risks with reinforcement learning agents operating in diverse environments [20]. Such contemporary AI agents could then be utilized as toy model for more complex anticipated AI safety issues [280]. More generally, this integration of contemporary AI as research object in the domain of AI safety is for instance reflected in the Asilomar AI principles which were crafted in 2017 [33] and endorsed by 1668 AI researchers and 3655 other individuals worldwide[4]. In addition, with the first of these 23 multifaceted principles being the general goal for AI research *"to create not undirected intelligence, but beneficial intelligence"*, it became obvious that AI development and safety required an extension beyond the classical boundaries of computer science. Similarly, the technology company Open AI crafted a publication in 2019 stating e.g. that *"properly aligning advanced AI systems with human values will require resolving many uncertainties related to the psychology of human rationality, emotion, and biases"* [242]. In the last years, the general requirement for a multidisciplinary approach to AI has as well been stressed e.g. in the domains of AI ethics [146], AI governance [127], responsible AI [336] and explainable AI [63].

Alternatively to conceiving AI safety as a twofold subject dealing with value alignment and control problem, it can be apprehended in risk-centered and cybersecurity-oriented

[4] These numbers correspond to the status displayed on the corresponding website as of April 2020. A full list of signatories is provided on `https://futureoflife.org/principles-signatories/`.



How and When did Type I system become Dangerous		Causes	
		On Purpose	By Mistake
Timing	Pre-Deployment	*a*	*c*
	Post-Deployment	*b*	*d*

Figure 1.1: Simplified overview of main Type I AI risks. Modified from [11].

taxonomic terms. In the style of a taxonomy of AI risks introduced by Yampolskiy [446], one could describe AI safety as *a discipline that addresses AI risks forming themselves at the pre- and post-deployment stage*. For clarity, consider the 4 main risks for Type I AI safety illustrated in the simplified Figure 1.1. First, risk *Ia* refers to intentional malevolent design by malicious human actors and could e.g. include the intentional crafting of an AI maximizing on harmful goals. Second, *Ib* is associated with intentional malicious attacks on deployed AI systems such as e.g. integrity-related attacks on AI sensors. Third, the risk *Ic* involves unintentional design-time mistakes with negative repercussions such as e.g. misspecified AI utility functions. Fourth, risk *Id* encodes unintentionally occurring operational failures such as e.g. misinterpretations of commands.

Originally, AI safety was mainly focused on unintentional problems (on risks *Ic* and *Id*) while neglecting security aspects of *intentional* malevolent design and malicious attacks [88, 338] (risk *Ia* and *Ib*). As described by Pistono and Yampolskiy [338], the cybersecurity paradigm aims by contrast at a balanced research comprising an exchange on both malicious exploits and safety measures for cyber-infrastructure – a balance analogously needed in AI safety. In accordance with this view, the thesis incorporates the eventuality of malicious actors in its analysis. In fact, it has been recently stressed in a security-relevant report that malevolent adversaries performing attacks on AI systems and malicious entities exploiting AI systems could have critical societal impacts in areas such as *"content filters, the military, law enforcement, traditionally human-based tasks being replaced by AI, and civil society"* [114]. For instance, by exploiting vulnerabilities of AI systems, content filters could be bypassed to spread unethical data, military systems could be systematically fooled, AI tools for law enforcement could be circumvented, attacks on intelligent systems could expose humans to dangers. Moreover, through malicious AI-based monitoring, parts of civil society could be specifically oppressed. For this reason, AI safety, AI ethics or AI governance approaches that do not address issues brought about by malicious actors may miss important facets of the security landscape[5].

[5] Appendix A provides a few examples of AI risk instantiations that already occurred in practice including some cases pertaining to malevolent actors (risks *Ia* and *Ib*).

Overall, this thesis provides the following 4 main contributions:

1. From Chapter 2 to 9, we paradigmatically analyze Type I AI safety in the context of the *meaningful control of intelligent systems* with a focus on the risks *Ib*, *Ic* and *Id* (whereby the risk *Ia* is addressed in the third contribution). For illustrative purposes, we consider the use case of autonomous vehicles.

2. In Chapter 10, we introduce the so-called *AI safety paradox* figuratively stating that value alignment and control represent conjugate requirements in AI safety. Consequently, we explain why a bifurcation of AI safety research into Type I and Type II AI safety is advisable and expound on why classical twofold AI safety cannot be solved.

3. In Chapter 12, we extract 10 (non-exhaustive) *transdisciplinary* and *hybrid cognitive-affective* strategical clusters out of our analysis tackling the Type I AI risks *Ia*, *Ib*, *Ic* and *Id* and complementarily also diverse instantiations of Type II AI risks. These 10 clusters implicitly touched upon within Chapter 2 to 11 range from large-scale conceptual AI governance to small-scale concrete AI engineering recommendations and comprise: *1) international (meta-)goals*, *2) transdisciplinary Type I/II AI safety and related education*, *3) socio-technological feedback-loop*, *4) integration of affective, dyadic and social information*, *5) security measures and ethical adversarial examples research*, *6) virtual reality frameworks*, *7) orthogonality-based disentanglement of responsibilities*, *8) augmented utilitarianism and ethical goal functions*, *9) AI self-awareness* and *10) artificial creativity augmentation research.*

4. In the final Chapter 13, we take the acquired insights and further transdisciplinary literature as basis to develop 3 concrete suggestions for future research directions within the two proposed branches of AI safety. These 3 research suggestions include: *1) Type I and Type II AI observatory*, *2) hybrid cognitive-affective defense methods for Type I AI* and finally *3) comparative transdisciplinary epistemology for Type I versus Type II systems.*

Outline

- Chapter 2 specifies requirements for the architecture of advanced Type I intelligent systems including the technical self-awareness property and collate a set of hybrid proactive Type I AI safety measures.

- Chapter 3 introduces the notion of ethical goal functions to control these Type I intelligent systems and elaborates on why it is possible to craft context-sensitive utility functions that are not touched by mathematically relevant consequentialist impossibility theorems.

- Chapter 4, we contextualize this proposed type of affective and perceiver-dependent goal functions within a novel non-normative scientifically grounded ethical framework termed augmented utilitarianism.

- Chapter 5 integrates the mentioned Type I architecture requirements for intelligent systems with the use of ethical goal functions within an AI governance framework denoted orthogonality-based disentanglement of responsibilities.

- Chapter 6 formalizes ethical goal functions from a cybernetics perspective and uses knowledge from psychology and cognitive neuroscience to inform this approach.

- Chapter 7 elaborates on virtual reality frameworks as support for human ethical debiasing and as counterfactual experiential testbed for the meaningful control of Type I intelligent systems as exemplarily applied to the use case of autonomous vehicles.

- Chapter 8 provides a short overview on opportunities that virtual reality frameworks offer to various Type I AI safety endeavors.

- Chapter 9 considers synergies between an international global humanitarian framework and sustainability issues in Type I AI safety utilizing the autonomous vehicle context as toy model.

- Chapter 10 introduces Type II AI safety and the AI safety paradox.

- Chapter 11 addresses the augmentation of both anthropic and artificial creativity as indirect approach to global challenges including AI safety.

- Chapter 12 concludes and explicitly enumerates the 10 hybrid cognitive-affective strategical clusters for AI safety implicitly identified throughout the thesis.

- Chapter 13 discusses the mentioned 3 suggestions for future research directions for Type I and Type II AI safety.

List of Publications

This thesis is based on the following set of 10 papers (enumerated in descending chronological order):

- N.-M. Aliman, P. Elands, W. Hürst, L. Kester, K. J. Thorissón, P. Werkhoven, R. Yampolskiy, and S. Ziesche. Error-Correction for AI Safety. In *International Conference on Artificial General Intelligence*, pages 12-22. Springer, 2020.

- N.-M. Aliman and L. Kester. Artificial Creativity Augmentation. In *International Conference on Artificial General Intelligence*, pages 23-33. Springer, 2020.

- N.-M. Aliman, L. Kester, P. Werkhoven, and S. Ziesche. Sustainable AI Safety? *Delphi – Interdisciplinary review of emerging technologies*, 2(4):226–233, 2020.

- N. Aliman, L. Kester and P. Werkhoven. XR for Augmented Utilitarianism. In *2019 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, pages 283-285. IEEE, 2019.

- N. Aliman and L. Kester. Extending socio-technological reality for ethics in artificial intelligent systems. In *2019 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, pages 275-282. IEEE, 2019.

- N. Aliman and L. Kester. Requisite Variety in Ethical Utility Functions for AI Value Alignment. In *Proceedings of the Workshop on Artificial Intelligence Safety 2019 co-located with the 28th International Joint Conference on Artificial Intelligence, AISafety@IJCAI 2019, Macao, China, August 11-12, 2019.*, 2019.

- N.-M. Aliman and L. Kester. Augmented Utilitarianism for AGI Safety. In *International Conference on Artificial General Intelligence*, pages 11-21. Springer, 2019.

- N.-M. Aliman and L. Kester. Transformative AI Governance and AI-Empowered Ethical Enhancement Through Preemptive Simulations. *Delphi – Interdisciplinary Review of Emerging Technologies*, 2(1):23–29, 2019.

- N.-M. Aliman, L. Kester, P. Werkhoven, and R. Yampolskiy. Orthogonality-Based Disentanglement of Responsibilities for Ethical Intelligent Systems. In *International Conference on Artificial General Intelligence*, pages 22-31. Springer, 2019.

- N.-M. Aliman and L. Kester. Hybrid Strategies Towards Safe "Self-Aware" Superintelligent Systems. In *International Conference on Artificial General Intelligence*, pages 1-11. Springer, 2018.

Chapter 2

Self-Awareness and Proactive AI Safety Measures

This chapter is based on a slightly modified form of the publication: N.-M. Aliman and L. Kester. Hybrid Strategies Towards Safe "Self-Aware" Superintelligent Systems. In *International Conference on Artificial General Intelligence*, pages 1-11. Springer, 2018. As the first author of the underlying paper, I had a vital contribution and it was solely my responsibility to write down the content and to perform an extensive literature research as well as in-depth analysis.

2.1 Introduction

Being a topic of major importance in AI safety research, AI alignment – which is often interchangeably used with the term of value alignment – has been analyzed from diverse points of views and incorporates a variety of research subareas many of which were reviewed by Taylor et al. [400]. Two highly relevant approaches in the realization of AI alignment the authors considered in this context are *value specification* and *error tolerance* which were both introduced by Soares and Fallenstein [392]. In order to do justice to these two distinct issues, Taylor et al. postulate that *"we can do research that makes it easier to specify our intended goals as objective functions"* concerning the first and *"we can do research aimed at designing AI systems that avoid large side effects and negative incentives, even in cases where the objective function is imperfectly aligned"* concerning the latter. We take these high-level considerations alongside additional multidisciplinary observations as point of departure and apply a more abstract and holistic analysis than many prior papers have utilized in this particular context to identify solution approaches. For instance, we see the need for "self-awareness" in intelligent systems for reasons such as safety, effectiveness, transparency or explainability just as such a functionality is required from the perspective of systems engineering for the effectiveness and safety of advanced models. Beyond that, we agree that methods inspired from cybersecurity practices [338] could provide a valuable support for AI safety including the safety of AGIs[1]. Furthermore, we also focus on the human factor in the AI development and suggest to make allowance for human cognitive constraints in AI safety while considering ethical aspects.

Outline In the next Section 2.2, we posit that a (yet to be defined) "self-awareness" functionality might beside other benefits account for an enhanced error tolerance within highly advanced to future human-level Type I AI models and might indirectly facilitate the value or goal specification process. Thereafter, in Section 2.3, we suggest that a self-aware advanced AI that should be deployed in a real-world environment will have to be supplemented by additional AI safety measures including for instance an AI red teaming[2] approach in order to maintain a high error tolerance level. In Section 2.4, we analyse how AI developers could proficiently face the problem of adequate value specification in the first place, which could interestingly imply the need for an enhancement of human "self-awareness" to a certain extent with respect to the goal to identify the values humans really intend on the one hand and regarding the aim to subsequently encode these values into prioritized goals a self-aware AI will have to adhere to on the other hand. Finally, in the last Section 2.5, we reflect upon this set of hybrid strategies as an interwoven entirety, consider its possible ethical implications and place it in the context of a hypothetically thereof emerging type of superintelligence.

2.2 Self-Awareness

While the notion of "self-awareness" which is often used in the context of concepts like "self-conciousness", "self-control" or "self-reference" is not in the focus of classical AI research, it is considered to be one of the key elements out of the crucial competency areas for human-level general intelligence according to many AGI researchers (as investigated by Adams et al. [3]) and the notion itself or related terms have been considered in some ways within various AI designs (e.g. in [39, 40, 192, 195, 318, 375, 404, 425]). However, the relevancy of AI self-awareness from the perspective of AI safety remains a poorly studied topic, even though the omission of such a functionality in an advanced AI architecture might lead to far-reaching implications in the future in regard to the safety of this system if deployed in a dynamic real-world environment. Given that a definition of this relatively abstract term is controversial and nontrivial, we will in the following first provide a simple technically oriented definition of AI self-awareness – for which we do not claim any higher

[1] Note that the term "AGI" as used in this chapter refers to future advanced *Type I* intelligent systems whose problem-solving ability exceeds or equates human problem-solving within the domains of interest.

[2] Red teaming refers to an attack simulation to identify vulnerabilities of systems in a given context.

suitability in general, but which is specifically conceptualized for our line of argument – and then subsequently elucidate the reasons for its crucial importance in AI safety frameworks.

The definition is inspired by systems engineering practices with applications to diverse types of dynamic systems as e.g. adapted by Kester et al. [261, 262] or van Foeken et al. [416] and is not restricted to the choice of any particular Type I AI architecture provided that the AI acts in a not further defined goal-oriented manner, possesses sensors and actuators as well as the ability to somehow communicate with human entities. For clarity, when we refer to an advanced AI exhibiting *self-awareness* in this work, we explicitly mean an advanced AI which is able to independently perform *self-assessment* and *self-management*, whereby self-assessment designates a set of processes enabling the AI to determine the performance of its various functions with respect to its goals (e.g. for associated physical instances, internal cognitive processes, own abilities, own resources,...) by itself and self-management the capability to adapt its behavior in the real-world on its own in order to reach its goals based on the information collected through self-assessment. In addition, the AI is presupposed to be able to communicate the insights obtained after having performed self-assessment and the choices made in the self-management step to specified human entities.

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

- *Transparency:* Through the ability of a self-aware AGI to allow important insights into its internal processes to its designers, it by design does not correspond to a "black-box" system as is the case for many contemporary AI architectures. The resulting transparency presents a valuable basis for effective AI safety measures.

- *Explainability:* Since the AGI performs self-management on the basis of a transparent self-assessment, its decision-making process can be independently documented and communicated, which might increase the possibility for humans to extract helpful explanations for the actions of the AI.

- *Trustworthiness:* An improved AGI explainability might increase its trustworthiness and acceptance from a human perspective, which might in turn offer more chances to test the self-aware AI in a greater variety of real-world environments and contexts.

- *Controllability:* Through the assumed communication ability of the AGI, a steady feedback loop between human entities and the AGI might lead to an improved human control offering many opportunities for testing and the possibility to proactively integrate more AI safety measures. More details on possible proactive measures are provided in the next Section 2.3.

- *Fast Adaptation:* Self-awareness allows for faster reactions and adaptations to changes in dynamic environments even in cases where human intervention might not be possible for temporal reasons which allows for an improved error tolerance and security. Unwanted scenarios might be more effectively avoided in the presence of negative feedback from the environment.

- *Cost-Effectiveness:* There is often a tradeoff between security and cost-effectiveness, however a self-aware system is inherently more cost-effective for instance due to the better traceability of its errors, the facilitated maintainability through the transparency of its decision-making processes or because the system can adapt itself to optimal working in any situation, while lacking any obvious mechanism which might in exchange lower its security level – by which a double advantage arises.

- *Extensibility*: Finally, a self-aware AGI could be extended to additionally for instance contain a coarse model of human cognition which could consider human deficiencies such as cognitive constraints, biases and so on. As a consequence, the AI could adapt the way it presents information to human entities and consider their specific constraints to maintain a certain level of explainability.

However, after having compiled possible advantages AI self-awareness could offer to Type I AI safety, it is important to note that up to now, it was not specified on what basis the goals of the self-aware goal-oriented AI are crafted in the first place. Moreover, the odds that a self-aware AI spawns many of the mentioned desirable properties are even largely dependent on the quality of the goals assigned to it and it is thus clear that self-awareness taken alone is far from representing a panacea for AI safety, since it does not per se solve the underlying goal alignment problem. Nonetheless, we argue that AI self-awareness represents a highly valuable basis for future-oriented AI safety measures due to the vitally important advantages it could bring forth if combined with appropriate goals. In addition, AI self-awareness might be able to itself facilitate the process of goal alignment through the interactive transparent framework suitable for tests in real-world environments it offers, whereby the selection of adequate goals clearly remains a highly debatable topic on its own. From our perspective, the therefore required goal function intrinsically reflecting desirable human values for a self-aware AI could be stipulated by humans which would be specifically trained in interaction with that AI and possibly ethically as well as cognitively enhanced on the basis of technological advances/scientific insights, since humanity at its current stage, seems to exhibit rather insufficient solutions for a thoughtful and safe future in conjunction with AIs – especially when it comes to the possible necessity for an unambiguous formulation of human goals. We will further address the motivations for human enhancement to provide assistance during this mentioned process of goal selection in Section 2.4.

2.3 Proactive AI Safety Measures

After having depicted possible benefits as well as still unanswered implications in the context of a self-aware Type I AI, we now focus on crucial AI safety measures which might be necessary in addition to avoid unintended harmful outcomes during the development phase and prevent risky scenarios after a subsequent deployment of such an advanced Type I AI architecture. While the suggested methods would undoubtedly not guarantee an absolutely risk-free AGI, their indispensability to at least obtain a well tested architecture built with a certain security awareness which particularly also takes the possibility of intentionally malevolent actors [338] into account, seems however to prohibit their omission. Beyond that, it seems imperative to incorporate a type of simulations of undesirable scenarios while developing an advanced Type I AI as a proactive rather than reactive approach, since the latter might be reckless given the extent of possible future consequences which could include a number of existential risks [83, 338, 401].

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

- *Development under adversarial assumptions:* Already during the A(G)I development phase, the developers should take into account the most important known types of e.g. integrity vulnerabilities that have been reported regarding other AIs in the past (this could include rather similar architectures, but importantly also cognitively less sophisticated AIs since it could represent a type of minimum requirement) and should not per default conjecture a benign environment. In a simplified scheme, assuming the development of an advanced AI starting nowadays, it should for instance among others be ascertained that none of the known adversarial methods to fool narrow AIs such as deep neural networks [326] would also lead to a defective information processing of security-relevant kind if correspondingly corrupted inputs are presented to the sensors of the AI at hand. Besides that, new types of A(G)I attacks and corresponding defense mechanisms should be actively ethically investigated. In this context a new subfield of study on "adversarial examples for Type I AGIs" appears recommendable. While adversarial examples for narrow AIs are for instance associated with definitions such as *"inputs to machine learning models that an attacker has intentionally designed to cause the model to make a mistake"*[3], a corresponding analogy could be derived for Type I AGIs. Ideally, the self-aware AI itself could be trained in identifying situations susceptible to involve particular known safety threats.

- *A(G)I red team:* As it is the case in the context of security systems, developers tend

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

to be biased towards emphasizing the robustness of their system and might additionally exhibit "blind spots" to existing vulnerabilities while implementing defense strategies [304], which is why realistic red team events offer an invaluable security tool in many cybersecurity frameworks [349, 351, 352]. Red teaming has recently as well be proposed by Brundage et al. [88] in the context of recommendations for an AI safety framework covering short-term issues for the next 5 years. Similarly, an external AI red team could in the long-term periodically perform real-world attack simulations after the deployment of an advanced AI, with the goal to identify certain types of possibly overlooked vulnerabilities to sophisticated attacks. The red team could for instance explicitly try to trigger unethical actions on the part of the AI by placing it in unknown or unusual contexts. In these settings, the blue team would correspond to the A(G)I developers which are responsible for the defense design within the AI architecture. Possibly, social engineering performed by the red team on the blue team could disclose biases underlying the AI training or its architecture and facilitate the crafting of specific targeted attacks. It is to be expected that such red team exercises will contribute to strengthen the robustness and possibly even enhance the cognitive abilities of the AI by providing the AI developers with comprehensive hints on how to enhance the defense designs which could for instance be of meta-cognitive nature. The ultimate objective would be to achieve a state from which on the self-aware Type I AI has learned to automatically and independently run self-tests simulating such systematical adversarial attacks.

- *Regular measurement of cognitive ability:* To maintain transparency and allow for a certain minimal monitoring of the Type I AI or AGI, it might be essential to be regularly aware of the level of cognitive ability it exhibits in order to customize the security measures. Besides classically proposed Turing tests, one further interesting type of test is the recently proposed "test for detecting qualia" introduced by Yampolskiy [447] based on visual illusions. However, instead of utilizing the described test measures to attempt to reveal the presence of qualia[4], we suppose that it could provide invaluable cues to detect higher cognitive abilities as exhibited by a Type I AI, since just like human misperceptions (including e.g. optical illusions) can for instance help to better understand the mechanisms underlying the perception of humans in cognitive science, so could the analysis of AI misperceptions analogously help to understand the internals of an advanced AI system. An automatic program could periodically test the AI and generate an alarm in the case of "cognitive anomalies" indicating an unusual increase of cognitive capacity. This regular test could also be implemented as a self-test mechanism within the self-aware Type I AI architecture itself. However, an explicit protective mechanism that prevents the AI or AGI from evolving any kind of harmful goals should be additionally designed in order to obviate any undesirable scenario. A related core idea to prevent an

[4] Simply put, qualia can be understood as conscious subjective experiences.

advanced AI from evolving a type of misaligned goal setting has been described by Goertzel [191] in the context of his suggestion for a specifically designed "AI Nanny" developed with a pre-defined set of goals and encompasses for instance *"a strong inhibition against modifying its [the AI Nanny's] preprogrammed goals"* or *"a strong inhibition against rapidly modifying its general intelligence"*.

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

2.4 Human Enhancement

Whereas in the context of the value alignment problem, the focus is often set on how future advanced Type I AIs could optimally learn values from human agents be it for instance by imitation or by predefined ethical goals, a jointly performed technology-supported learning approach for human agents to enhance their cognitive abilities and ethical frameworks in order to be able to develop improved capabilities qualifying them to more competently deal with this highly relevant problem in the first place, remains an under-explored topic. Given the large array of human deficiencies including for instance cognitive biases [455], unintentional unethical behavior [386] or limitations of human information processing which could be considered as major handicaps in succeeding to solve the AI alignment problem, the approach to extend the abilities of humans in charge of developing an ethical AI or AGI by science and technology emerges as an auspicious strategy, however certainly not without reservations.

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

- *Ethical enhancement:* One prominent subproblem of goal alignment can be simply described as to make the AI learn human goals [401]. For this purpose, humans obviously need to be first aware of the values they really intend to implement in order to encode them as a factual set of prioritized goals within an advanced Type I AI model. Similarly, as stated in [30], humans need to become better "ethical regulators" (e.g. of themselves and of AIs) in an era which will be more and more shaped by AI. This task might inter alia require a better type of "self-assessment" on the part of humans – especially with regard to their own concrete ethical preferences, abilities and constraints. To improve the required human ethical self-assessment for the development of safe AIs, developers should consider a dynamic multifarious science-based ethical framework which could for instance encompass debiasing training [309] as well as methods from behavioral ethics [148] and could in the future even include a type of Type I AI-assisted debiasing training where the same self-aware Type I AI which is periodically checked for safety could e.g. act as "teacher" in game settings providing a personalized feedback to its developers which could be expanded to a testing of acquired ethically relevant skills. Additionally, the group formation of the AI or AGI developers itself should ideally reflect a synergetic heterogeneity of worldviews to fend off inequality and unnecessary biases at the core of the goal selection process.

- *Cognitive enhancement:* Some decades ago, the cybernetics pioneer Ross Ashby expressed the following train of thought [31] : *"[...] it is not impossible that what is commonly referred to as "intellectual power" may be equivalent to "power of appropriate selection". [...] If this is so, and as we know that power of selection can be amplified, it seems to follow that intellectual power, like physical power, can be amplified."* Even if this statement might still reflect a controversial issue and human enhancement technologies are still in their infancy, expected progresses in areas such as nanorobotics, bionics, biotechnology, brain-computer interface research or the newly arisen field of cyborg intelligence integrating *"the best of both machine and biological intelligences"* [388] might lead to considerably extended possibilities for cognitive enhancement in the foreseeable future. Transferring the term used in Ashby's statement to a different context, we argue that (possibly Type I AI-assisted) methods to increase the human "power of appropriate *goal* selection" within the framework of A(G)I development given the ethical values agreed upon while supported by preceding ethical enhancement procedures, represent an essential future research direction to be pursued for AI safety reasons. For this purpose, one could first experimentally improve on presently clearly not sufficient enhancement concepts such as mental training, human-machine interface tools, neurofeedback, non-invasive brain stimulation methods, multi-mind brain-computer interfaces for decision-making or nootropics. Later on, a reasonable priority for a self-aware Type I AI might even be to generate methods facilitating human cognitive enhancement and develop concepts where if procurable the Type I AI augments rather than surrogates human entities initiating a bidirectional learning framework. Besides that, the group composition of A(G)I developers should ideally promote multidisciplinarity in order to reduce the occurrences of AI safety relevant blind spots in the development phase and should comprise numerous partcipants with diverse research backgrounds.

While it should be clear that human enhancement pathways cannot guarantee the prevention of an occurring unethical Type I AI [10], not to perform human enhancement does not guarantee it either. Furthermore, the abstention from ethical human enhancement also does not necessarily prevent the performance of unethical human enhancement by malevolent actors at a later stage. Therefore, we argue that the early practice of human enhancement for ethical purposes like the improvement of the value specification process for AI alignment, might increase the odds of a resulting ethical AGI and could even in the long-term facilitate the detection of potential unethical AGI development or unethical human enhancement through the bundled cognitive and ethical abilities that could emerge out of the suggested bidirectional framework of mutual enhancement.

2.5 Conclusion and Future Prospects

We postulated that Type I AI self-awareness represents a highly valuable functionality from the perspective of AI safety as it might be helpful for the error tolerance subtask of AI alignment as well as indirectly for value specification and provides many advantages such as transparency or explainability. We then introduced a number of proactive AI safety measures including A(G)I red teaming which could be necessary in addition to the self-awareness functionality to maintain security and which might be beneficial for the error tolerance subproblem. We set forth that the described framework alone might not be sufficient due to the ethical and cognitive constraints AI developers exhibit as human beings and proposed a jointly performed inter alia AI-assisted ethical as well as cognitive enhancement procedure to support the goal selection process. We do not claim that the described hybrid framework represents a complete approach warranting the safety of the Type I AI or of a therefrom emerging superintelligence, but argue that it might underpin the importance of a multidisciplinary approach to AI safety and motivate a new useful holistic perspective on the complex problem of AI alignment which might in turn shape future developments towards a beneficial form of Type I superintelligence (i.e. an AI with problem solving abilities exceeding the human baseline across the most relevant domains of interest). Finally, we stress that possible future research on self-aware Type I AIs as well as research on ethical and cognitive enhancement for AI safety should not be reserved to stakeholders like corporations, the military or a presumed elite group of AI or AGI developers, but be instead performed open-source and shared across diverse communities for the benefit of mankind. Moreover, a science-based debate on the implications of disruptive technological advancements [342] should be encouraged and existential risks through Type I superintelligence should be thoroughly analyzed – especially regarding scenarios implying the presence of malicious actors [10, 338].

2.6 Contextualization

In this chapter, a preliminary approach to the value alignment problem (value specification and error tolerance) has been discussed. Among others, self-awareness and proactive AI safety measures have been proposed to increase error tolerance and indirectly support the value specification process. While the focus in this chapter was generally on highly advanced Type I AI, the self-awareness concept has been described as a challenge for the computer science community and been identified as requirement for the meaningful control of any intelligent system in a follow-up work by Werkhoven et al. [428] – making self-awareness already essential for present-day projects on intelligent systems. However, for a full account of the value specification subtask, the briefly mentioned concept of a goal function encoding human ethical values and legal constraints is necessary in addition. This type of utility function is denoted *ethical goal function* in the following.

The complex endeavor of equipping intelligent systems with an appropriate ethical goal function can be understood as an AI governance task and additionally requires a scientifically grounded approach to morality with elements from psychology and cognitive science as will become apparent in the next chapters. For instance, it was initially assumed that such ethical goal functions would need to correspond to a consequentialist and utilitarian framework given that a mathematical utility function was required to explicitly quantify ethical and legal constraints. Thereby, the need for classical utilitarian utility functions $U(s')$ assigning utilities to the outcomes of actions widely used in the AI field [363] for Markov decision process related "rational" agents and borrowed from economics was an implicit assumption. The reason being that back then, alternative theoretical motivations for cardinal utility functions were missing and utilitarian utility functions were seen as sole option as a consequence.

However, as will be expounded in the next Chapter 3, Eckersley [152] showed in the meantime that such utilitarian objective functions/utility functions face certain impossibility theorems if applied in contexts where human lives and well-being measures are at stake (which could be the case for certain types of intelligent systems). Simply put, classical utilitarian utility functions were shown to be unable to capture human ethical intuitions which seemed to entail that ethical goal functions for meaningful control in sensible safety-critical contexts are per definitionem impossible. Would that conclusion be true, it would raise multiple security-relevant open questions on the feasibility of meaningful control of intelligent systems. However, the following Chapter 3 identifies a solution to the puzzle by pointing out a possibility (and the linked necessary conditions) to formulate *context-sensitive* and *affective* ethical goal functions that would not be touched by the mentioned impossibility theorems. For this purpose, we analyze the underlying philosophical and mathematical considerations while incorporating aspects from psychology and elements from a branch of mathematics denoted order theory[5]. This novel approach *differing from classical utilitarianism* paves the way for a reformulation of the concept of utility functions. Moreover, this view incorporating affective elements is more scientifically plausible as touched upon in Chapter 4 and 6. Beyond that, Chapter 7 conveys the necessity of conceiving of human enhancement measures for ethical debiasing in AI safety as *cognitive-affective.*

[5] Order theory [324] is a subfield of mathematics dealing with binary relations. The link to utilitarian utility functions $U(s')$ can be illustrated by a simplified example. Consider a set of outcomes (of actions) $S = \{s'_1, s'_2\}$ and a binary relation $\geq$ determining the subjective ethical desirability of an outcome such that $s'_1 > s'_2$ iff s'_1 is ethically more desirable than s'_2. In the case $s'_1 > s'_2$ holds, a corresponding cardinal utility function $U(s')$ formulated at the level of the outcomes s' of actions would encode that $U(s'_1) > U(s'_2)$.

Chapter 3

Transformative AI Governance and Preemptive Simulations for Ethical Goal Functions

This chapter is based on a slightly modified form of the publication: N.-M. Aliman and L. Kester. Transformative AI Governance and AI-Empowered Ethical Enhancement Through Preemptive Simulations. *Delphi – Interdisciplinary Review of Emerging Technologies*, 2(1):23–29, 2019. As the first author of the underlying paper, I had a vital contribution and it was solely my responsibility to write down the content and to perform an extensive literature research as well as in-depth analysis.

3.1 Introduction

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 chapter, we evince why for the purpose of effective AI governance, humans have to quantitatively specify their ethical conceptions within a utility-based framework. Thereby, the implementation of advanced Type I 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 a general argument against the feasibility of such a utility-based 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.

3.2 Utility-Based 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 intelligent systems: 1) prohibitive, 2) self-regulative, 3) deontological and 4) utility-based 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 deontological method 3) whose goal is to embed ethical values in AI systems via deontological rules and the utility-based 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 utilized 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 realize 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 [82] 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 utility-based 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” [428]). Conversely, for the utility-based 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 [12]) 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 [428], a problem which a utility-based 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 utility-based solution would only require a centralized 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 utility-based 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 realize efficient AI governance measures.

3.3 Dynamical Ethical Enhancement

The realization of a utility-based approach to AI governance utilizing ethical goal functions should be considered as a dynamical process in which these functions are steadily reviewed and updated. Given a domain, the legislative could provide anethical 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 maximizing 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 after every update of ethical goal functions. 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" [428] 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 Type I AI safety, this socio-technological feedback-loop might immanently contribute to tackle 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, this 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.

3.4 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 [152]. For this reason, it is of critical importance to make sure that ethical goal functions are able to safely encode desirable conceptions on population ethics which are not in conflict with those of the society that crafted it. Population ethics [214] 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"* [29] 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[1]. 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 [328].

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 [29] 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 [214]. 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[2]. 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 general 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, a utility-based decision-making should not be necessarily regarded as a remote, detached and passive endeavor, 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 e.g. known that *"anticipatory emotions arise in reaction to mental discrete images of the outcome of a decision"* [54] and that this mental simulation phenomenon termed "conceptual consumption" [187] provides a basis for decision-making. Moreover, to consider the thereby experienced utility in this immediate

[1] However, note that in a few paragraphs it will become apparent that this type of utility assignment can be too simplistic.

[2] Note, that this finding does not only apply to consequentialist frameworks, since every classical normative moral theory needs a population axiology which is why e.g. deontological analogues for impossibility theorems are similarly conceivable, see [214].

hedonic sense is much closer to the original idea of "utility" as introduced by Jeremy Bentham [68]. Therefore, we argue that a society willing to perform a utility assignment with the goal to achieve a contextualized 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 e.g. 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 *mental-state-dependence* of population axiology. Due to this dependence, it is cogitable that *different mental states could potentially lead to different population axiologies* i.e. 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 *mental-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 *mental-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 valenced mental simulation of this scenario. This new total order placing population Z before population A would however appear natural to most people[3]. This circumstance can be explained by the introduced mental-state-dependence of population axiology. *Without a mental context, the utility of a state has no meaning in ethics.* To sum up, coming back to the realization 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 – *as long as their nature is inherently mental-state-dependent*[4]. 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.

3.5 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. Admittedly, so far, we did not concretize 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 [254] 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

[3] Put very simply, if the utility assigning population M has less than 8 billion people (such as the world today), choosing a policy-making measure p to get to either A (a population with 10 billion members and very high positive welfare) or Z (a population of 10.000 billion members with much lower and only barely acceptable but still positive welfare), leads to the selection of A. The reason being that it appears that $U(A) > U(Z)$ since in this context, the transition (M, p, A) seems ethically preferable and because (M, p, Z) seems “repugnant” in comparison. If however M consists of 10.000 billion members like Z, the transition (M, p, A) suddenly seems repugnant instead and $U(Z) > U(A)$ simply because to get from M to A might signify that p implies e.g. a genocide reducing 10.000 billion people to 10 billion people.

[4] The next Chapter 4 provides more details on the novel framework required to encode such mental-state-dependent i.e. affective and context-sensitive utility functions.

represents a more realistic model of hedonic experience which is directly linked to human wellbeing/ 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"* [254]. 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 [253]. 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"* [254]. Since this circumstance might lead to ethical goal functions that do not maximize 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 utilization 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 Section 3.3), 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 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.

3.6 Conclusion

If it holds that objective happiness represents a suitable bottom-up approach to well-being/happiness[5], then an ideal strategy to promote human well-being, would be to implement ethical intelligent systems able to maximize on the aggregated simulated future instant utility (i.e. 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 e.g. the PERMA model of positive psychology [382] 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 summarized that the presented utility-based 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 mental-state-dependent ethical goal functions.

[5] From the perspective of constructionist accounts in psychology such concepts would be instead understood as constructions and a valence-based instant utility could be categorized as one of the affective ingredients of every construction. This view is further considered in Chapter 6.

3.7 Contextualization

This chapter expounded why a Type I AI governance strategy based on utility functions is not categorically impossible if those functions additionally encode information about the mental simulations underlying moral judgements. Such simulations do not only consider the outcome of actions but also *the context* within which these outcomes occurred. This could for instance include the nature of the actions and the initial state from which on actions are performed. It has been argued in this chapter that by allowing a context-sensitive and mental-state-dependent perspective beyond purely behavioristic considerations, ethical goal functions could overcome the constraints of classical utilitarian objective functions linked to consequentialism. However, it seems premature to regard the problem as solved without further specifying the fundamental difference between this novel type of context-sensitive utility functions and their classical counterpart. Furthermore, such a clarification would facilitate a subsequent formalization necessary for guidelines to obtain an appropriate machine-readable format. In the next Chapter 4, we elaborate further on this novel framework underlying context-sensitive and affective utility functions which we denote *augmented utilitarianism.* Importantly, it is crucial to note that augmented utilitarianism does *not* represent a subtype of utilitarianism/consequentialism since it allows to consider simulations encoding *agent, action and outcome* as opposed to these views. Moreover, it does *not* represent a prescriptive normative framework. In short, the goal of augmented utilitarianism is *not* the formulation of ethical imperatives motivating "what humans ought to do", but it represents a *descriptive* and explanatory scientific endeavor aiming at aggregating information relevant to human moral judgements in a more adequate form than prevailing utility functions (and in a manner that is untouched by the mentioned impossibility theorems) to facilitate meaningful control. One guiding question is rather for instance: *"how to formulate ethical goal functions that do not violate human ethical intuitions?"*. While so far, the field of AI safety addressed ethical frameworks for AI systems mainly from the perspective of philosophy, this new endeavor necessitates considerations from scientific disciplines such as cognitive science and psychology. In our view, philosophical thought experiments often offer excellent opportunities for illustrations in AI safety, however a grounding in science is necessitated to accurately *model* human ethical conceptions and try to capture meaning in ethics pertaining to intelligent systems – a requirement for meaningful control. Thus, the next chapter uses philosophical questions from AI safety related to moral judgements as starting point but then theoretically motivates the need for augmented utilitarianism from a psychological and cognitive sciences perspective. In a later Chapter 6, we address more practical details and mathematical formalization aspects for ethical goal functions crafted with augmented utilitarianism.

Chapter 4

Augmented Utilitarianism for AI Safety

This chapter is based on a slightly modified form of the publication: N.-M. Aliman and L. Kester. Augmented Utilitarianism for AGI Safety. In *International Conference on Artificial General Intelligence*, pages 11-21. Springer, 2019. As the first author of the underlying paper, I had a vital contribution and it was solely my responsibility to write down the content and to perform an extensive literature research as well as in-depth analysis.

4.1 Introduction

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 the "literalness"[301, 445] or also "perverse instantiation" [83, 444] 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 AI safety and AI 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* [14, 428] 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 AIs [159, 160] (including classical ethical frameworks like consequentialism or encompassing methods based on uncertain objectives and moral uncertainty [79, 152]), most approaches do not provide a fundamental solution to the underlying problem which wrongly appears to be solely of philosophical nature. According to Goertzel [193], *"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.

4.2 Deconstructing Perverse Instantiation

Using the generic notation $< FinalGoal > : < PerverseInstantiation >$, we enumerate 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 [83])
2. "Make us happy" : "Implant electrodes into the pleasure centers of our brains" (example by Bostrom [83])
3. "Making all people happy" : "Killing all people [...] as with zero people around all of them are happy" (example by Yampolskiy [445])
4. "Making all people happy" : "Forced lobotomies for every man, woman and child [...]" (example by Yampolskiy [445])

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* [14] 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 [337, 383], hedonic psychology [253] 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) [382]. Another known psychological measure for well-being is subjective well-being [290] (SWB) which is composed of frequent positive affect, infrequent negative affect and life satisfaction [91, 138]. In both cases, happiness only represents a subelement of the respective well-being construct. Similarly, as stated by Diener and Bieswas-Diener [139], *"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 [14]. 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 [152].

Population ethics [214] 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 [214], 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 [328]. In this context, Arrhenius [29] 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 in Chapter 3 (in [14]), 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 [14]. 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 AIs that are implemented in the form of expected utility maximizers mostly operating in a consequentialist fashion.

[1] Importantly, this also applies to non-consequentialist frameworks such as deontological ethics [214].

4.3 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 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 AI 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 [67], 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* [14] 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 chapter, we exemplarily consider well-being as arbitrary reasonable high level final goal candidate which is e.g. already reflected in the UN Sustainable Developmental Goals (SDGs) [461] and is in the spirit of positive computing [93]. Besides SWB [290] and PERMA [382], multiple measures of well-being exist in psychology with focus on different well-being factors. For instance, the concept of objective happiness [253] 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 [258, 266, 297]. (Recently, related AI research aiming at inducing self-transcendent states for users has been considered by among others Mossbridge and Goertzel [310].) For a review on relevant well-being factors that might be pivotal for a dedicated positive computing, see Calvo and Peters [91].

- *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 [54]) has been termed conceptual consumption [187] and plays a role in decision-making. Similarly, according to Johnson [248] *"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* [170, 256] 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 [48], the brain constructs mental states based on *"sensations from the world, sensations from the body, and prior experience"* [325]. 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 [38]. Ethical frameworks for AI 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[2] 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 [348]) 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 [69, 254].

[2] How to possibly aggregate utility assignments of different perceivers to a societal-level perspective is briefly addressed in Chapter 6.

Since it is impossible to directly extract the instant utility (the basic building block of experienced utility [254]) 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 [254]. Therefore, it has been suggested in Chapter 3 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 [428], this technique might offer a powerful preemptive tool for AI 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 [93, 417]. Interestingly, the experience of individualized and tailored simulations itself might provide an alternative simulation-based solution to the value alignment problem [449].

- *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 [14]. 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 AI 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.

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**

Table 4.1: Decision-making focuses within different possible ethical frameworks for AI safety. "S&T" denotes a foreseen augmentation of the ethical decision making process by science and technology including AI 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).

- *Amalgamation of diverse perspectives*: Finally, we postulate that AU[3], despite its intrinsically different motivation as a socio-technological ethical framework for AI 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 4.1 provides an overview on the different decision-making focuses used in relevant known ethical frameworks that might be seen as candidates for AI safety.

4.4 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 AI safety. While CU and other classical ethical frameworks if used for AI 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 [14] (which differs from using util-

[3] AU is not be to confused with agent-relative consequentialism which – as opposed to AU – is a normative agent-based framework, does not foresee a grounding in science and seems to assume a "pretheoretical grasp" [376] of its "better-than-relative-to" relation

ity functions implicitly learned by AI agents or AIs learning moral conceptions from data such as e.g. in [348]). 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 [193] 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 [12, 428] as described in Chapter 2. 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 AI safety research [12] on *ethical adversarial examples* – a conceivable type of integrity attacks on the AI sensors having ethical consequences might be important to study in future work to complement the use of safe utility functions.

4.5 Contextualization

In Chapter 3, we introduced a systems-engineering oriented solution for AI governance refered to as orthogonality-based disentanglement of responsibilities. In this context, a missing piece was a suitable ethical framework to craft ethical goal functions. Now that AU has been identified to be able to fill that gap for the support of a meaningful control of intelligent systems, it seems expedient to provide a preliminary integration of the elements suggested so far in the thesis and to touch upon open questions. Given an AU-based ethical goal function for a specific domain, a compatible safe type of AI architecture capable of self-management and self-assessment is required to reliably perform actions maximizing on that goal function. In the next Chapter 5, we recapitulate the motivation for the mentioned disentanglement of responsibilities, provide more details on how to implement self-management and self-assessment in AI systems, present clarifying illustrations and formulate tentative future-oriented AI governance recommendations on this basis. We also elaborate further on the notion of a socio-technological feedback-loop which also emphasizes the need to modify ethical goal functions with time in order to facilitate corrigibility.

Chapter 5

Orthogonality-Based Disentanglement of Responsibilities for Ethical Intelligent Systems

This chapter is based on a slightly modified form of the publication: N.-M. Aliman, L. Kester, P. Werkhoven, and R. Yampolskiy. Orthogonality-Based Disentanglement of Responsibilities for Ethical Intelligent Systems. In *International Conference on Artificial General Intelligence*, pages 22-31. Springer, 2019. As the first author of the underlying paper, I had a vital contribution and it was solely my responsibility to write down the content and to perform an extensive literature research as well as in-depth analysis.

5.1 Introduction

In the current both safety-critical and ethically relevant international debate on how to achieve a meaningful control of advanced intelligent systems that comply with human values [365], diverse solution approaches have been proposed that fundamentally differ in the way they would affect the future development of A(G)I research. In a nutshell, one could identify a set of four main clusters of conceptually different solution approaches for which one could advocate for by distinguishing between 1) *prohibitive*, 2) *self-regulative*, 3) *deontological* and 4) *utility-based* methods. While the prohibitive approach aims at restricting or even banning the development of highly sophisticated AI until problems related to control and value alignment are solved in the first place, it seems highly unlikely to be put into practice especially in its most extreme forms and it is therefore not further considered in this chapter. By contrast, option 2) implies the assumption that certain mechanisms (for instance specific market mechanisms or mechanisms inherent to certain types of A(G)I architectures) could allow for a more or less automatically emerging stability or desirability of the behavior as exhibited by intelligent systems. Furthermore, solution 3) classically considers the direct hard-coding of ethical values into AI systems for instance by encoding deontological values at design time [414], while in the case of the utility-based approach 4), one mostly foresees a human-defined utility function [456] quantitatively encoding human values. This debate – especially on whether to prefer the solution approach 3) or 4) – is often strongly imprinted by particularly difficult to solve philosophical issues and the AI-related responsibilities of different involved stakeholders such as users, programmers, manufacturers and legislators appears to be only vaguely and therefore insufficiently definable. Against this backdrop, the need for a practicable technically oriented and at the same time forward-looking solution appears to be of urgent importance for a responsible future planning of a hybrid society in close conjunction with advanced AI systems.

5.2 Disentanglement Of Responsibilities

For reasons of safety, security, controllability, accountability and reliability, it can be assumed that it is in the interest of a democratic society to achieve a transparent division of responsibilities for the deployment of intelligent systems in diverse application areas. Thereby, the systems should act in accordance with ethical and legal specifications as formulated by the legislative power and allow for traceability in order to facilitate an assignment of responsibility by the judicial power. Consequently, we argue that the self-regulative solution 2) can be ruled out since it would lead to a heterogeneous set of different ethical frameworks implemented within different types of intelligent systems yielding highly complex entanglements especially with regard to responsibility assignments (e.g. among manufacturers, programmers, users and operators). Furthermore, as the problem solving ability of the intelligent systems increases, the severity of possible unintended effects, malicious attacks [88] or the development of intentionally crafted unethical systems [338] which could even induce existential risks seems to prohibit a laissez-faire approach. Thus, the remaining options are the deontological approach 3) and the utility-based solution 4) since both could be in theory implemented within a framework separating the responsibilities as described.

According to the orthogonality thesis by Bostrom [83], *"intelligence and final goals are orthogonal axes along which possible agents can freely vary"*. Though, the thesis is not uncontroversial for reasons comprising the fact that it does not address probabilities as postulated by Goertzel [194]. However, for the purpose of our specific argument, it is not necessary to consider the soundness of the thesis, since we only presuppose that "there exists a type of AI architecture for which final goals and intelligence are orthogonal" which is self-evident considering utility maximizers [27] as classical examples epistomizing solution 4). From this, it trivially follows that formulating a goal function for a utility maximizer and designing the architecture of this agent are separable tasks. Building on that, we argue that the already existing practice of the legislative power having a say on the *what* goals to achieve as long as societal impacts are concerned and the manufacturers implementing the *how* in various contexts can be adapted to goal-oriented utility maximizers (albeit with certain reservations particularly on the nature of the architecture used) and can thus be pursued as postulated by Werkhoven et al. [428].

Apart from that, it is undoubtedly possible to think of a similar disentanglement of responsibilities in accordance with a solution of the type 3). However, for mostly technical reasons we will now illustrate, we do not consider a deontological framework in which lawful and ethical behavior is encoded for instance in ontologies [236] or directly in natural language as possible instantiation of our orthogonality-based disentanglement approach. First, the attempt to formulate deontological rules for every possible situation in a complex unpredictable real-world environment ultimately leads to a state-action space explosion [428] (it is thereby obvious that law does not represent a complete framework). To be able to handle the complexity of such environments and the complexity of internal states, the intelligent system needs to be run-time adaptive which cannot be achieved by using static rules. Second, since law is formulated in natural language which is inherently highly ambiguous at multiple linguistic levels, the intelligent system would have to either make sense of the legal material using error-prone Natural Language Processing techniques or in the case of the ontology-based approach, the programmers/manufacturers would have to first interpret law before encoding it which induces uncertainty and violates the desired disentanglement of responsibilities. Third, law leaves many legal interpretations open and entails tradeoffs and dilemmas that an intelligent system might encounter and would need to address leading to an unspecified assignment of responsibilities. Fourth, an update of laws will require a costly and laborious update of designs for every manufacturer. Fifth, a deontological approach with fixed rules cannot easily directly endorse a process in which progresses in AI could be efficiently used to transform society in a highly beneficial way enabling humans to overcome their cognitive and evolutionary biases and creating new possibilities to improve the foundations of society.

Having expounded why the deontological solution approach 3) is inappropriate for the central problem of disentangling responsibilities for the deployment of intelligent systems, we now elucidate how a properly designed solution 4) is able to avoid all mentioned disadvantages associated with solution 3). First, it might be possible to realize run-time adaptivity within utility maximizers by equipping them with a "self-awareness" functionality [12] (self-assessment, self-management and the ability to deliver explanations for actions to human entities) which we outline in Section 5.4. Moreover, deontological elements could be used as constraints on the utility function of such utility maximizers in order to selectively restrict the action or the state space. Second, by quantifying law within a publicly available ethical goal function as addressed in the next Section 5.3, one achieves an increased level of transparency. Third, through a utility function approach tradeoffs and dilemmas are more easily and comprehensibly solved. Thereby, for safety reasons, the utility functions can and should include context-sensitive and perceiver-dependent elements as integrated e.g. in augmented utilitarianism [13]. Fourth, updates of law are solely reflected in the ethical goal functions which leads to a more flexible and controllable task. Fifth, the use of such an ethical goal function opens up the opportunity for a society to actively perform an enhancement of ethical abilities which we depict in Section 5.5.

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

A first step of crafting ethical goal functions could be for instance to start with the mapping of each relevant application domain of law to a specific utility function which quantifies the expected utility of the possible transitions of the world. For this purpose, the legislative has for instance to define the relevant components of each goal function and assign weights to each component, decide which parameters to consider for each component and identify possible underlying correlations. (It is thinkable that specific stakeholders might then while applying the goal function to their particular area of application, craft a lower-level customized mission goal function [153] for their specific mission goals which would however have to be compliant with the ethical goal function provided by the legislative.) The implementation of this strategy will require a relatively broad multidisciplinary knowledge by policy-makers or might require the practical collaboration with trained multidisciplinary researchers with expertise in e.g. AI and systems engineering.

One important feature of the proposed framework is the requirement of transparent human-readable goal functions that can be inspected by anyone which substantially facilitates accountability. In order to obtain a specification of a human-oriented goal function, different methods have been proposed including inverse reinforcement learning (IRL) [160] and reward modeling [279]. However, the IRL method comes with the main drawback of yielding ambiguous reward functions that could explain the observed behavior and within reward modeling, a black-box model is trained by a user in order to act as reward function for a reinforcement learning agent which violates both the transparency requirement of our approach and the disentanglement of responsibilities since it is the user that trains the reward model (and not a representation of society).

However, it is important to note, that as implicit so far, the goal functions would be rather specified based on *what humans want* and not necessarily on what humans *should* want from a scientific perspective, since it is known that humans exhibit biases for instance inherent to their neural systems [267], due to their evolutionary past of survival in small groups [428] or through ethical blindspots [386] which represent serious constraints to their ethical views. On these grounds, the framework described in this chapter is intended to be of transformative and dynamical nature and might enable the legislative to receive a quantitatively defined feedback from the environment, which in turn might foster the human-made evidence-based adjustment of the explicitly formulated ethical goal functions towards more scientifically sound assumptions.

Beyond that, as postulated by Harris [225], a *science* of morality which might enable humans to identify the peaks on the "moral landscape" which he described as *"a [hypothetical] space of real and potential outcomes whose peaks correspond to the heights of potential well-being and whose valleys represent the deepest possible suffering"* could represent a feasible general approach to solve moral issues. In the light of the aforesaid, one could attempt to in the long-term pursue research that facilitates the design of a scientifically grounded universal ethical goal function whose local optima will ideally be conceptually equivalent to the peaks of this hypothetical moral landscape potentially reflecting what humans *should* want. Another interesting point of departure to be mentioned in this context, has been introduced by Ziesche [461] who describes how the UN sustainable development goals already representing an international consensus and containing values such as well-being could be quantified to start to practically tackle the value alignment problem (more details in this regard are analyzed in Chapter 9).

Note that Yudkowsky' s early idea of a coherent extrapolated volition [454] in the context of friendly AI which envisaged an AI maximizing the utility based on an extrapolation of what we *would* want *"if we knew more, thought faster, were more the people we wished we were, had grown up farther together"* while being relatively close to it, is though subtly different from our described concept of what we *should* want based on a scientifically grounded ethical goal function, since an improvement of our problem solving ability does not necessarily improve our ethical abilities nor does *"the people we wished we were"* necessarily corresponds to a more ethical version of ourselves on average. Moreover, there is no reason to assume that human values would necessarily converge to ethical values if they *"had grown up farther together"*. However, as will be introduced in Section 5.5, our method of utilizing ethical goal functions aims at actively grounding the implementation of ethics in a transformative socio-technological feedback-loop for which the legislative provides the seed.

5.4 “Self-Aware” Utility Maximizer

After having commented on the procedure of crafting ethical goal functions, we now describe a class of architectures able to yield controllable utility maximizers that strictly comply with a generic goal function specified by humans. In the following, we explain how a top-down analysis leads to an exemplary technically feasible and minimalistic instance of this class. Note that when we refer to an intelligent system in the following, we specifically mean a system able to independently perform the OODA-loop (Observe, Orient, Decide, Act). One can further decompose the system into four distinct cognitive functions: sensors, orienter, decision maker and actuators according to these four subcomponents respectively. In a first step, we assume that the utility maximizer cannot be based on a subsymbolic learning paradigm *alone* (such as Deep Learning (DL)), since desirable reactions to all possible situations an intelligent system could encounter in complex real-world environments cannot be learned in reasonable time with finite computational resources. Thus, we postulate in a second step that a certain level of abstraction is required which can be achieved by combining a symbolic reasoning component with a perception exploiting the advantages of learning algorithms resulting in a “hybrid architecture”. However, this hybrid intelligent system needs to be as well-equipped with a self-model to face the possible complexity of its internal processes without which the system would be confronted with similar problems caused by the inability to anticipate reactions to all possible internal states. In a third step, we argue that the requirement for a self-awareness capability [12] comprising self-assessment and self-management as well as the ability to provide counterfactual explanations for actions to human entities appears essential for instance for reasons such as the necessity of constructing solutions in real-time that have not been learned before including sensor management [261], adaptivity in the case of communication to other intelligent systems [262] and for explainability purposes. Apart from this, the view expressed by Thorissón [404] that *“self-modeling is a necessary part of any intelligent being”* which similarly considers the importance of feedback-loops relating the actions of a system to the context of its own internal processes could be a further argument supporting the relevance of self-awareness.

Taking these requirements into account, one feasible instance of the described class of hybrid self-aware utility maximizers could integrate DL algorithms – presently representing relatively accurate machine learning models especially in the vision domain – as sensors at the subsymbolic level able to output classification results that can be further processed by the orienter component yielding a symbolic representation of the situation and the internal processes. As decision maker one could envisage a utility-based reasoning/planning (and not learning) process such as e.g. with (partially observable) Markov decision processes (MDP) equipped with the ethical goal function as specified by the legislative, a causal model of the world and of the system itself. The decision maker would map sym-

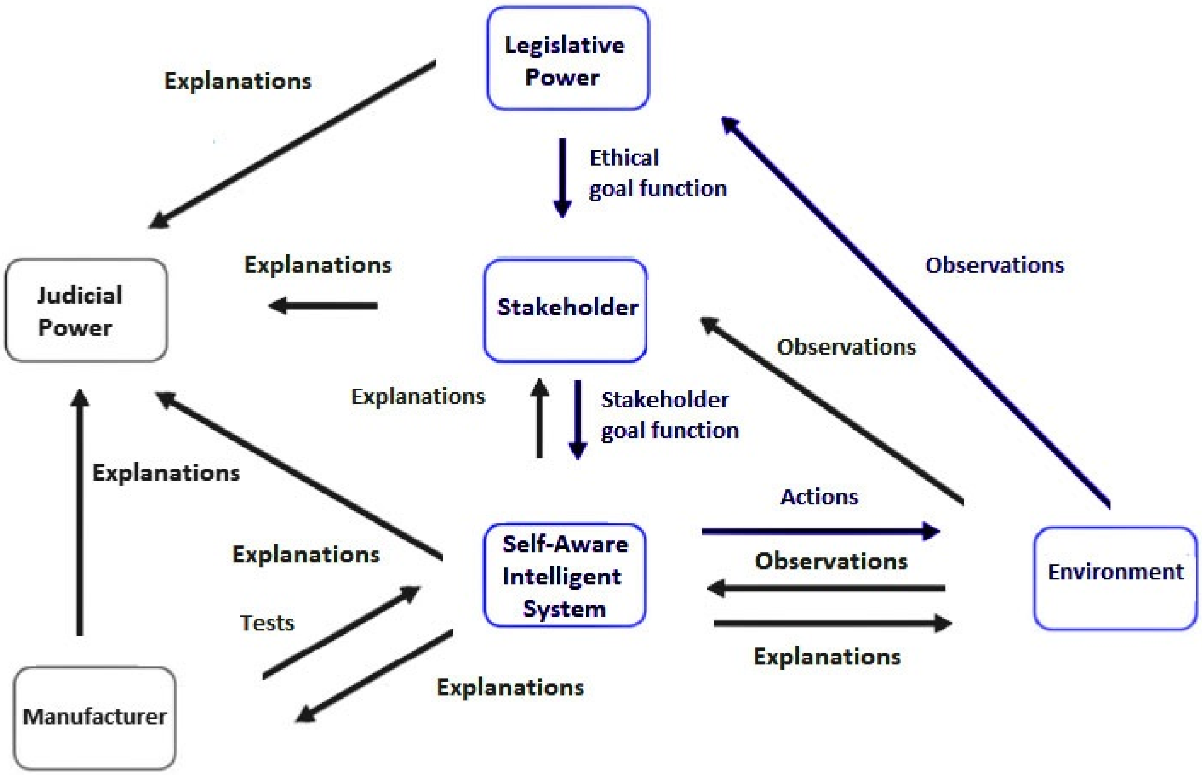


Figure 5.1: Simplified illustration and contextualization of a socio-technological feedback-loop (highlighted in blue) implementing the orthogonality-based disentanglement approach for a generic stakeholder domain.

bolically encoded situations and internal processes to actions maximizing on expected utility with respect to the ethical goal function that are finally executed by the actuators either on the environment or on the system itself. In this framework, explanations could be delivered at the symbolic level. Concerning the input-to-output mappings of the DL sensors, one possibility could be to strive to monitor the related uncertainty by means of self-management which will have to be reflected in the goal function.

5.5 Socio-Technological Feedback-Loop

Having discussed how a disentanglement of societal responsibilities for the deployment of intelligent systems could be achieved, introduced the notion of an ethical goal function and described the corresponding requirements an intelligent system might need to fulfill in order to comply with such a function, we illustrate and contextualize the composite construction of a consequently resulting socio-technological feedback-loop in Figure 5.1. At the pre-deployment stage, the manufacturer is responsible for verification and validation practices including the conduct of system tests demonstrating the ability of the intelligent system to adhere to the ethical goal function. At post-deployment stages, the judicial power determines for instance whether the different agents acted in compliance with an ethical goal function given a set of explanations. Concerning the main socio-technological feedback-loop, its key characteristic lies in the fact that it would enable the legislative to dynamically perform revisions of an ethical goal function based on its *quantifiable* impacts on the environment and that it could serve as powerful policy-making tool. Thereby, this feature is paired with the peculiarity that the nature of the environment is not restricted to solely encompass real-world frameworks. More precisely, one could for instance distinguish between three different variations thereof enumerated in an order of potentially increasing speed of formulating/testing hereto related policy-making measures that might be substantiated in an ethical goal function: 1) classical *real-world environments*, 2) specifically crafted and constrained *synthetic environments* and 3) *simulation environments*.

Since the design of an appropriate ethical goal function represents a highly complex task and the necessary time window to collect evidence on its societal impacts in real-world settings on a large-scale might often represent an undesirable complication, policy experimentation on a small-scale in restricted synthetic environments relating the ethical goal function to specific impacts might represent a complementary measure. However, an even more efficient solution allowing for faster decision-making is the "policy by simulation" approach [428] in which human expert knowledge can be extended by AI systems within simulation environments. In doing so, AI might finally assist humans in developing more ethical AI systems while ultimately enhancing human ethical frameworks by relating the mathematic formulation of an ethical goal function to its direct impacts on the (simulated) environment making possible answers to the crucial question on "what humans *should* want" graspable and beyond that, potentially a direct object of scientific investigation. Finally, the proposed orthogonality-based disentanglement of responsibilities could provide a new perspective for the AI coordination subtask in Type I AI safety– the non-trivial issue of making sure that global AI research is dovetailed in such a way that no entity actually implements an unethical and unsafe AI – e.g. by offering a starting point for considerations towards an international consensus on the principle of using publicly accessible ethical goal functions that can be easily inspected by the public and international actors. This method might reduce the AI race to the problem-solving ability dimension while at the same time providing incentives for demonstrably ethical and transparent frameworks tightly coupled to an ethical enhancement of partaking societies. Given that the law already represents a public matter, it does thereby not seem to represent an exceedingly disruptive step to advocate for public ethical goal functions.

5.6 Conclusion and Future Prospects

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

However, certain crucial safety and security challenges remain to be separately addressed and should be taken into consideration in future work. First, self-improvement within an intelligent system could for instance be implemented by an online learning process or by reconfigurability through run-time adaptivity. While it is reasonable to avoid self-improvement by learning during the deployment of the system in order to limit safety risks, future work will need to examine the possibility of verification methods for self-improvement by reconfigurability at run-time. Second, while the self-awareness functionality facilitates (self-)testing mechanisms, extended research on the controllability of specific test procedures in synthetic testing environments will be required. Third, a turn-off action could be seen as a primitive form of self-management in the context of tasks where the performance of the system superseded human performance. However, the possibility to turn-off the system for security reasons by specified human entities should always be given. Fourth, for the purpose of malevolence prevention, it is important to rigorously consider proactive security measures such as A(G)I Red-Teaming at the post-deployment stage and research on adversarial attacks on the sensors [12, 407] of the self-aware intelligent system. Fifth, a blockchain approach to ensure the security and transparency of the goal functions themselves and all updates on these functions might be recommendable. Crucially, in order to avoid formulations of an ethical goal function with safety-critical side effects for human entities (including implications related to impossibility theorems for consequentialist frameworks [152]), it is recommendable to assign a type of perceiver-dependent and context-sensitive utility to simulations of situations instead of only to the future outcome of actions [14, 13]. In the long-term, we believe that scientific research with the goal to integrate the first-person perspective of society on perceived well-being within an ethical goal function at the core of the presented socio-technological feedback-loop might represent one substantial element needed to promote human flourishing in the most efficient possible way aided by the problem solving ability of AI.

5.7 Contextualization

In this chapter, it has been theoretically motivated that ethical goal functions for the control of Type I intelligent systems should be iteratively crafted by a representation of society (e.g. the legislative). This strategy could also enable a broader acceptance and could represent a human-centered approach. However, what remains unclear is how to practically and mathematically formalize these non-consequentialist utility functions to make them fit-for-purpose. 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, the next Chapter 6 performs a transdisciplinary analysis applying a security mindset to the issue and summarizing variety-relevant background knowledge from cognitive neuroscience and psychology. We complement this information by linking it to augmented utilitarianism as a suitable ethical framework. Based on that, the next chapter proposes first practical guidelines for the design of approximate ethical goal functions that might better capture the variety of human moral judgements.

Chapter 6

Requisite Variety in Ethical Utility Functions for AI Value Alignment

This chapter is based on a slightly modified form of the publication: N. Aliman and L. Kester. Requisite Variety in Ethical Utility Functions for AI Value Alignment. In *Proceedings of the Workshop on Artificial Intelligence Safety 2019 co-located with the 28th International Joint Conference on Artificial Intelligence, AISafety@IJCAI 2019, Macao, China, August 11-12, 2019.*, 2019. As the first author of the underlying paper, I had a vital contribution and it was solely my responsibility to write down the content and to perform an extensive literature research as well as in-depth analysis.

6.1 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 [223, 392, 456] (a review is provided in [400]). Moreover, the need to investigate value alignment has been included in the Asilomar AI Principles [33] with a worldwide support of researchers from the field. While value alignment has often been tackled using reinforcement learning [2] (and also reward modeling [279]) or inverse reinforcement learning [1] 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* [14, 428] (see Chapter 3). 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 [116] 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, Type I 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) [47, 263]. Thereby, psychological functions such as cognition, emotion or moral judgements are closely linked to the predictive regulation of physiological needs of the body [263] 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 6.2 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 6.3, 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* [13] 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 [14]. Finally, in Section 6.4, we conclude and specify open challenges providing incentives for future work.

6.2 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 [198], this variant has been denoted *ethical adversarial examples* [13]) which could lead to an unethical behavior even if the utility function would have been aligned with human ethical

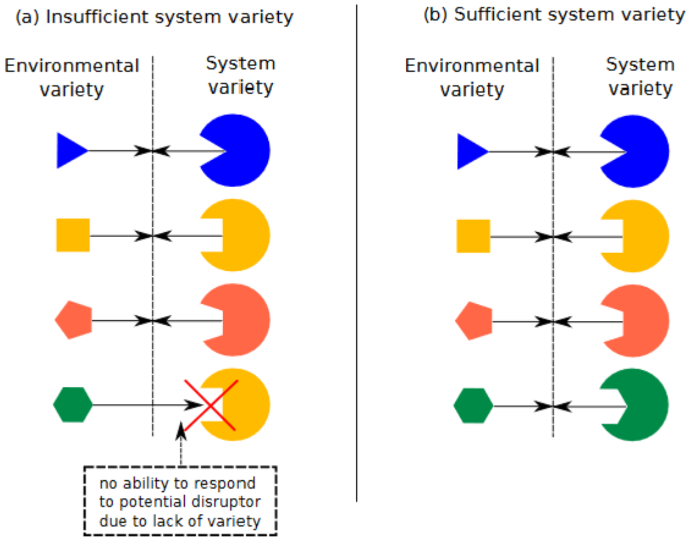

(a) Insufficient system variety
Environmental variety
System variety
no ability to respond to potential disruptor due to lack of variety
(b) Sufficient system variety
Environmental variety
System variety

Figure 6.1: Intuitive illustration for the law of requisite variety. Taken from [320].

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 [199] 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 model[1] that given a sample, outputs a value determining the perceived ethical desirability 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 [31] 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 6.1 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

[1] A conceptually similar separation of objective function model and optimizing agent has been recently performed for reward modeling [279]

possible undesirable outcomes, is by increasing the own variety – which can be achieved by encoding more relevant knowledge.

6.2.1 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"* [291] 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 [47, 302]. 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 [48, 52].

Emotion and cognition do not represent a dichotomy leading to a conflict in moral judgements [230]. 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 [53] 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"* [237]. The mental phenomenon of actively dynamically simulating different alternative scenarios (including anticipatory emotions) has also been termed conceptual consumption [187] and plays a role in decision-making and moral reasoning. While emotions are more akin to discrete constructions, core affect allows a low-dimensional experience of interoceptive sensations (statistical regularities of the internal milieu) [48] and is a continuous property of conciousness with the dimensions of valence (pleasantness/unpleasantness) and arousal (activation/deactivation) [263]. It has been argued that core affect provides a basis for moral judgements in which different events are qualitatively compared to each other [92]. Like other constructed mental states, moral judgements involve domain-general brain processes which very 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) [325]. 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 [95]) 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"* [47] – 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 [47]. 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 [4]. 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.

6.2.2 Variety through “Dyadicness”

The psychological theory of dyadic morality [373] 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 $iA \overset{\mathrm{d}}{\rightarrow} vP$. 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 [371]. The combination of these basic constituents is suggested to lead to the emergence of a rich diversity of moral judgements [211]. *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 [374]. 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"* [373].

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* [13] (abbreviated with AU in the following) which was introduced in Chapter 4. Reconsidering the dyadic morality template $iA \overset{\mathrm{d}}{\to} vP$, it seems that in order to better capture the variety of human morality, utility functions – now transferring it to the perspective of Type I AI systems – would need to be at least formulated at the abstraction level of a *perceiver-dependent* evaluation of a transition $s \overset{\mathrm{a}}{\to} s'$ 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 $U_x((s, a, s'), t)$.

6.3 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 6.2, 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 [14]) 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.

6.3.1 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 $s \xrightarrow{\mathrm{a}} s'$ 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 [13]):

$$U_x(s, a, s') = v_x(m_x((s, a, s'), B_x, P_x, E_x)) \tag{6.1}$$

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 [13]. 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* [14] (see Chapter 3). The use of consequentialist utility functions affected by the impossibility theorems of Arrhenius [29] has been justifiably identified by Eckersley [152] 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 [421] 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. [106] 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 [14]). 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 6.3.2, 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.

6.3.2 Approximation, Aggregation and Validation

So far, it has been stated throughout the chapter 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 deliberate 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 [382]) with Type I 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 $A \xrightarrow{\text{w}} P$ in analogy to the cognitive template of dyadic morality. Thereby, $A \xrightarrow{\text{w}} P$ 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 [382]. 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 $A \xrightarrow{\text{w}} P$ 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 $U_x(s, a, s')$. (Note that an alternative high-level final goal compatible with the abstraction level of $U_x(s, a, s')$ could as well be harm minimization. A conceivable perceiver-dependent cognitive template could be formulated as $A \xrightarrow{h_{min}} P$ encoding a (human or artificial) agent A performing an act h_{min} that causes the least possible harm to a human patient P. Overall, this more pragmatic goal might even appear preferable in practice, since sustainable safety and well-being cannot be guaranteed as we touch upon in Chapter 9.)

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 $A \xrightarrow{\text{w}} P$ 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 (e.g. the temporal development of societal satisfaction or the temporal development of perceived harm reduction with AI systems in a certain domain), 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 (representation of) 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[2] given a domain:

$$U_{Total}(s, a, s') = \sum_{n=1}^{N} \sum_{i=1}^{j} w_i^n f_{fi}^n(C_i) \tag{6.2}$$

with N standing for the number of participating entities in society, $C_i = (p_{i1}, p_{i2}, ..., p_{im})$ 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 $f = \{f_1, f_2, ..., f_f\}$ 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 $O = \{C_1, C_2, ...\}$ 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 $A \xrightarrow{\text{w}} P$ or $A \xrightarrow{h_{min}} P$). 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 $U_{Total}(s, a, s')$).

[2] Note that the following solely represents one possible examplary aggregation strategy and that multiple alternative valid methods are thinkable (see e.g. the options specified by Masthoff in [298]).

Finally, it is important to note that the societal ethical goal function specified in (2) will need to be updated (and evalutated) at regular intervals due to the mental-state-dependency of utility entailing time-dependency [13] (see Chapter 4). This leads to the necessity of a socio-technological feedback-loop which might concurrently offer the possibility of a *dynamical ethical enhancement* [14, 428]. Pre-deployment, one could in the future attempt a validation via selected preemptive simulations [14] in which (a representation 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* [14] denoted $U_{TotalAS}$ being a potential constituent of future well-being. Thereby, $U_{TotalAS}$ refers to the instant utility [254] 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:

$$U_{TotalAS}(s, a, s') \approx \sum_{n=1}^{N} \int_{t_0}^{T} I_n(t)dt \tag{6.3}$$

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 establishing a link to harm minimization) that has to be a priori determined.

6.4 Conclusion and Future Work

In this chapter, we motivated the need in Type I 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 cognitive 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"* [13]. 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 [428] 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 [14].

6.5 Contextualization

This chapter revealed among others the indispensability of affective and dyadic parameters in ethical frameworks for intelligent systems and the importance of perceived harm in human moral judgements as described in the theory of dyadic morality. To make the numerous preceding theoretical reflections more graspable, the next Chapter 7 analyzes recent experimental work in virtual reality (VR) studies which especially attempted to assess human moral judgements as applied to ethical decision-making in the context of autonomous vehicles (AVs) – an exemplary Type I intelligent system. Interesting questions arising here are for instance *"how to design VR studies that can do justice to the variety of perceived harm predicted by dyadic morality?"* and *"how to scientifically debias misinformed moral cognition (while respecting moral pluralism) with the help of VR?"*. Answering these questions might be helpful to identify appropriate parameters for ethical goal functions. Hence, in the next Chapter 7, we postulate that for the complex task of societal relevance pertaining to goal specification in both AI ethics and AI safety, VR and also augmented reality (AR) represent valuable tools whose utilization facilitates the extension of socio-technological reality by offering a rich *counterfactual experiential testbed* for enhanced ethical decision-making. For this purpose, we use the example of AVs to elaborate on how especially VR could provide a twofold structured augmentation for the governance of artificial intelligent systems by enhancing society with regard to ethical self-assessment and ethical debiasing. Thereby, we extend existing literature by tailored recommendations based on insights from cognitive neuroscience and psychology to solve inconclusive open issues related to past VR experiments involving ethically relevant dilemmas in AV contexts.

Chapter 7

Extending Socio-technological Reality for Ethics in Intelligent Systems as exemplified by the Autonomous Vehicle Case

This chapter is based on a slightly modified form of the publication: N. Aliman and L. Kester. Extending socio-technological reality for ethics in artificial intelligent systems. In *2019 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*. IEEE, 2019. As the first author of the underlying paper, I had a vital contribution and it was solely my responsibility to write down the content and to perform an extensive literature research as well as in-depth analysis.

7.1 Introduction

For the governance of artificial intelligent systems which is a field of interest within both AI safety and AI ethics at an international level [365], it becomes crucial to design an appropriate goal specification framework able to encode the ethical and legal requirements within a given societal context. In this regard, different solutions have been proposed ranging from rule-based frameworks to methods based on updatable context-sensitive and perceiver-dependent ethical utility functions formulated at the societal level called *ethical goal functions* [16] (see Chapter 5). However, for any case in a given domain in which a society is supposed to contribute to implement a framework of AI governance, a process of *ethical self-assessment* attempting to unambiguously provide answers to the question on *what society wants* arises as necessity. In order to identify strategies for legislations and regulations, it seems vital to proactively carefully analyze moral judgements and moral actions during human-machine interactions. In the context of autonomous vehicles (AVs), VR [434] or AR settings might allow for qualitatively superior estimations in comparison to rather remote online questionnaires [38] while concurrently facilitating safety-relevant experiments which would be proscribed in real-world environments [70]. How VR has already been used to assist in ethical self-assessment is reviewed and discussed in the following Section 7.2.

While this procedure might enable a first assessment of the status quo of moral views and the identification of ethical parameters, the findings are at times contradictory and society is not limited to directly pass on these conceptions which are potentially distorted by ethically relevant cognitive biases to the AVs without more ado. First, the self-assessment itself may be prone to biases and misconceptions – a circumstance that can be improved using scientific models of morality [9, 373] as basis for better suited VR and AR studies. Second, it appears recommendable to schedule a scientifically grounded process of *ethical debiasing* in order to filter out *what society should want* given what society wants. For this purpose, the utilization of VR or AR has been suggested for the experience of counterfactual preemptive simulations [14] in which people can scientifically calibrate their beliefs and as a result their moral judgements due to discernable differences in experience[1]. Ways in which the described ethical debiasing via VR/AR experiences could take place in the future and how this could be linked to a dynamical enhancement within a socio-technological feedback-loop [16] in the spirit of Positive Computing [93] are elucidated in Section 7.3.

7.2 Ethical Self-Assessment

In order to identify what society wants in an AV context, different studies have used especially VR settings implementing either different ethical dilemmas inspired by classical trolley problems [255, 399, 434] or conceptualized for more practical scenarios [238, 282]. In our view, both types of studies appear valuable due to the fact that while ethical dilemma scenarios have been criticised to be unrealistic because AV scenarios are more complex [322], one could argue that if these simplicistic scenarios can already not be resolved, it seems a rather pessimistic outlook for real-world AV cases implying decision-making under uncertainty. Another line of thought is to assume that ethical dilemmas like trolley problems are inherently insoluble [238] and thus, the only engineering task that can be fulfilled is to avoid lethal outcomes at any cost. However, it is known from cybersecurity that *"there is no such thing as a 100% secure system"* [451] which also

[1] Note that the conceptual design of such VR or AR experiences themselves would require the integration of knowledge from multiple relevant scientific fields to allow for a targeted identification of ethically relevant parameters in the first place. Exemplary recommendations are provided in Subsection 7.3.2.

generalizes to AI safety. Therefore, any proactive approach could eventually fail and it is important to plan for such eventualities in order to enable risk assessment and cost-benefit analysis. Importantly, while encountering ethical dilemmas with lethal consequences in real-life AV settings might represent a relatively rare case, the design of suitable solutions could provide invaluable insights for intelligent systems in critical domains where the lives of people and their well-being are inherently part of the decision-making process. Thinkable relevant application areas include justice, medicine and military [152].

Furthermore, one could argue that the perception of AVs does not allow a distinction of people of the same quality as supposed in many ethical dilemma scenarios. However, this argument seems not to indicate a fundamental limitation, since a more sophisticated perception with a better integration of data is rather a matter of time, knowledge and legal framework and does not represent a categorical impossibility. Thus, both classical ethical dilemmas and more real-world oriented VR and AR studies appear to be of interest for AI ethics and AI safety considerations that should support ethical self-assessment even if practical cases are likely to be more often of relevance in daily life [353]. In the following, we review some relevant results reported from a non-exhaustive set of VR experiments of both types which yields a variety of heterogeneous and partly even contradictory interpretations.

7.2.1 Ethical Dilemmas

In a versatile VR experiment, Wilson and Theodorou [434] analyzed the moral judgements of participants immersed in the perspective of AV passengers occupying the driver seat in different settings. Thereby, the decision maker component of the AV was faced with the inevitable choice to crash within one of two different humanoid non-playable characters. The authors studied i.a. the impact of the varying perception of the agent on the reactions of the participants. In all three conditions, the AV was in fact controlled by an AI, however in one condition the participants were informed that a human was remotely controlling the car, while in both remaining conditions, the participants were informed about the artificial control but with a split leading to a transparent vs. a non-transparent setting. Among others, the authors reported an important difference in the acceptance of the decision-making of the agent by the participants dependent on whether they assume it to be of artificial or human nature. The participants were more likely to exhibit forgiveness for the actions of the “human” and non-transparent AI agents than for the transparent AI, which was however perceived as more predictable. Moreover, in case of decision-making based on the social value of the accident victims (for instance using a distinction by profession, gender and body size), the participants tended to blame the transparent AI much more than the agent perceived as human. Further, the participants strongly favored random decisions rather than such based on social value to which they vehemently opposed. It is important to note that while the graphics were slightly limited in their naturalness, every crash was accompanied with realistic screams from real recordings. Overall, this study indicated a difference in moral judgments related to the perspective of the participants, since questionnaires framing comparable AV dilemmas from a non-immersive bird's eye view [38] led to different results with participants basing their moral judgments i.a. on differences in social value.

Other VR studies in the context of decision-making in AVs led to diverging conclusions. For instance, Kallioinen et al. [255] analyzed moral judgements in VR from various perspectives (passenger, pedestrian and observer) with human or artificial drivers and came to the conclusion that humans and AVs were largely judged similarly. Thereby, the only important difference was related to the stronger expectation for harm minimization in the AV case. Generally, participants favored to spare the lifes of a smaller group of child pedestrians over a larger number of lifes of adult pedestrians. Moreover, they exhibited a tendency to spare the life of pedestrians over car passengers given an equal number of individuals, however with the restriction that from the passenger perspective, participants *"were less likely to accept the car driving off a cliff in order to save pedestrians"*. The authors link this fact to a supposed self-preservation effect. In this study, the self-reported confidence of moral judgements was additionally registered with the lowest confidence scores assigned to adult vs. child pedestrian scenarios from the more detached observer perspective. An important detail is that the authors state to have omitted realistic animations and sound effects at collision time in order to *"avoid unnecessary distress"*.

In a further VR study [399] analyzing moral behavior instead of retrospective judgement in a forced choice decision with vehicles involving a variety of obstacles ranging from objects over animals to humans, the participants were controlling a car from a driver's seat and had to indicate via keystrocke which of two obstacles to target by lane switch. In this experiment, the authors similarly abstained from any sound and animation effect associated with the collision itself. In addition, the experiment was subdivided in a fast and a slow condition of a 1 second and a 4 seconds time window for decision-making respectively. The authors argue that a simple value-of-life model approximates the taken moral decisions appropriately but observe inconsistencies in moral behavior under time pressure. They postulate that value-of-life models should be used for real-world applications in the AV context and are of importance for manufacturers and lawmakers. Another virtual car driving study [163] concludes that human decision-making in similar ethical dilemmas can be captured by a utilitarian model in favor of the quantitative greater good sparing the lives of as many virtual avatars as possible (irrespective of whether this includes a self-sacrifice of the own avatar). In addition, the study identifies a modulation of the decision process with respect to the age parameter associated with the victim avatars. The younger the avatars were, the less likely it was that the participants decided to sacrifice the lives of these entities. This type of "age bias" was as well observed in a similar VR study [398] involving ethical dilemmas with AVs where the authors however additionally report a “gender bias” in favor of female avatars and a slight “omission bias” describing the preference to hit targets by inaction instead of actively switching the lane. An interesting detail is that in this study, the authors also omitted animations and sound effects at collision time and provided a time window of 4 seconds for each decision.

In a different cross-cultural VR experiment which was not overtly framed within an ethical dilemma paradigm, but which involved an immersive game-like situation of a race-course with simulated car accidents [250], the impact of personality traits on the nature of the decision-making in such emergency situations has been analyzed. In this context, Ju et al. designed the simulated inevitable accidents occuring during a VR race with recorded lap time such that they reflect ethically relevant decision-making. The main goal of the experiment was formulated as the completion of a car race, however three humanoid avatars appeared on the street with different aversive behaviors towards the approach of the car. The participants had the possibility to either brake/try to avoid the situation, run over the avatars or drive off a cliff resulting in a self-sacrifice. The time window for decision-making amounted to less than 2 seconds. Thereby, the control of the car was designed such that the brakes were too unsensitive to avoid the collision. Consequently, the race ended either just before the collision or once the participant opted to drive off the cliff. Under these conditions, only one participant opted for a self-sacrifice, while the choices of all participants were splitted in two distinct groups of those that did not ignore the avatars (60 participants) and those that did (30 participants). The latter group exhibited significantly higher psychopathy-related traits assessed using a self-reported psychopathic scale, an empathy scale and an interpersonal reactivity index. These findings implying a strong aversion against self-sacrifice stand in contradiction to the VR studies presuming an impersonal utilitarian decision-making in car accident scenarios. Simultaneously, it is in line with the results of studies according to which people are highly reluctant to accept utilitarian AVs that would involve a self-sacrifice by the passengers, even though they would want others to possess such vehicles – leading to a social dilemma [81]. Moreover, most VR experiments for AV decision-making so far did not consider the importance of personality traits in the evaluation procedure which seems however to be essential.

7.2.2 More Practical Scenarios

Realistic AV scenarios exhibit richer safety-relevant features than many of the presented ethical dilemma scenarios which share similar structures as known from classical trolley problems. For instance, while the decision-making is performed from the perspective of an individual, AVs will require a societal-level approach. Moreover, the VR experiments do not necessarily reflect a situation involving moral and in particular legal responsibility [322] due to the missing penalization and the risk that participants might not grasp the seriousness of the task [250]. When analyzing the results obtained in these studies it becomes apparant that potential conflicts between widely hold ethical conceptions and the state-of-the-art legal frameworks for AVs might arise. An example for this are the German guidelines binding for AVs stating that *"personal features, such as age, should not be taken into consideration in unavoidable accident situations"* [255]. Against this backdrop, it appears vital to jointly analyze both ethical and legal dimensions within immersive experiences for a more differentiated assessment in order to identify future integrated societal strategies for the deployment of AVs and more generally ethical intelligent systems.

In [282], Li et. al present a VR study in which the moral action dilemmas include the parameter of compliance with (Chinese) traffic laws by the victims and further also motorcyclists and car drivers. The authors found that next to sparing the lifes of the greater number of avatars, people tend to for instance explicitly attribute a higher value to the life of pedestrians complying with the law and thus perceived as innocent in comparison to pedestrians acting unlawfully by crossing the street at a red traffic light. In general, the participants tended to protect the life of pedestrians more often (especially children) than the better equipped motorcyclists or car drivers. However, when experiencing pedestrians acting unlawfully, the number of people increased who would spare lawful motorcyclists or car drivers when compared to a scenario in which no avatar violated the traffic laws. Interestingly, the study comprised two assessments of each situation. In a first step, the participants were immersed in the perspective of the driver and had to act under time pressure, while a second confrontation with the situation from a more remote bystander view allowed a reconsideration of the selected choice before the final decision in the form of vehicle dynamics was saved. Within the study, the deliberation taking place during the second confrontation led some participants to alter their decision to spare avatars acting in compliance with the traffic laws in order to both minimize their own accident liability and to save the life of avatars perceived as innocent. The authors argue that law enforcement is required to constrain the AV deployment since *"no one want to see an AV running into them at random"* [282] and suggest to protect innocent groups of people who abide by the law rather than blindly employing a utilitarian decision-making framework. For this purpose, they propose an obligatory ethics setting for AV deployment and an integration of a quantification of expected crush injury severity.

7.3 Ethical Debiasing

After having reviewed a number of VR studies that have been performed for a societal-level ethical assessment aimed at facilitating the governance of AV systems relevant for AI safety and AI ethics, it becomes clear that the highly heterogenous, equivocal and partly contradictory nature of the results exhibits the need for a deeper analysis. In this chapter, we postulate that integrating a modern scientific understanding of the nature of human morality based e.g. on insights from cognitive neuroscience and psychology could help to clarify the underlying issues and may provide hints for the design of better targeted VR and AR studies in order to debias the ethical self-assessment procedure itself by avoiding misconceptions i.a. based on outdated psychological models and unsound assumptions. Furthermore, we agree with Zuromski et al. stating that [463] *"VR technologies are a form of extended mind. In fact, they are "mental institutions" and they enable various cognitive processes [...] which are in fact new tools of social cognition"*. Hence, we postulate that besides representing an invaluable mental extension to identify *what a society wants* (ethical self-assessment), VR technologies might be in addition of essential importance to better assess *what a society should want* (ethical debiasing) [14] – seemingly the conditio sine qua non for a responsible deployment of artificial intelligent systems in real-world environments.

It appears too restrictive to limit the breadth of human ethical conceptions in AV scenarios to a dichotomy between utilitarian and deontological decision-making often misleadingly framed as a "rational" vs. an "emotional" view [399, 434]. From a virtual ethical perspective [321], it has been argued that *"we should try to design and program cars in ways that help to make people act carefully and responsibly when they use self-driving cars"*. Thereby, it is conceivable that VR experiments could act as a catalyst for AVs to become "moral technologies" [7] which could facilitate an ethical enhancement of society and has been stated to be able *"to bring out virtues in people"* [321]. In the following, we elaborate on why an ethical framework for AVs would need to allow a coalescence of the diverse apparently conflicting ethical conceptions as found in the VR studies in addition to legal constraints in order to function as an adequate embodied scientific model of human intuitions and how this might affect the design of future suitable VR experiences. Therafter, we briefly elucidate how to possibly use VR and also AR to further refine and debias the resulting goal specification.

7.3.1 Scientific Ethical Self-Assessment

It has been argued from a cybernetic point of view that in order not to violate human ethical intuitions, the framework utilized to govern an intelligent system should be a model of these intuitions whereby the best models pertaining to human mental constructions are scientific ones [9]. From the perspective of AI safety, the governance of AI systems requires solving the value alignment subtask which aims at implementing AI systems such that they are aligned with human ethical values. Hence, for a responsible AI governance it is important to scientifically consider the nature of human morality and to overcome popular but unfounded conceptions related to the human brain supposing for instance a battle between emotion and cognition at the core of its functioning which is often nourished by conceptions like the “triune” brain representing a neuroscientifically untenable assumption [47, 302]. This idea related to a presumed dichotomy between a “rational” neocortical and an “emotional” limbic thinking mode has been termed a *“cherished narrative in Western mental philosophy”* and a fiction with implications even in economical settings [168]. Traces of this type of mind-body dualism can be found in multiple approaches to morality and may be reflected in certain scientific interpretations of moral psychology. In fact, mental states associated with both emotion and cognition are constructed via domain-general processes in a brain for the purpose of allostasis (anticipating the needs of an organism before they arise) [47]. The experience of instances of concepts as emotion on the one hand and cognition on the other hand differs in the focus of attention of the brain [53, 237]. Thereby, limbic cortices are not reactive as assumed, but predictive and control the body as well as attention mechanisms among others [48, 52]. Moreover, next to environmental information and past experiences, the construction of mental states in general (such as thoughts, moral judgements, or emotions) involves the consideration of the interoceptive sensory array (statistical regularities of the internal milieu) as additional necessary element for a situated conceptualization for the purpose of allostasis. Interoceptive sensations are made available to human conciousness via the low-dimensional continuous experience of core affect with the dimensions valence (pleasantness/unpleasantness) and arousal (activation/deactivation) [48]. Importantly, *“no decision or action can be free of interoception and affect”* [47] which is why also moral judgements and moral actions always involve affective elements to a certain extent even in cases where humans associate them with the adjectives “rational” or “cold” [9].

Hence, while it is tempting to contrast utilitarian decision-making in the VR scenarios with the more deontologically seeming decisions by using a “rational” vs. an “emotional” account of the situations, it is important to step back and consider alternative consistent scientific explanations in order to be able to better assess what a society wants. In light of this, note that an early psychological theory termed dual-process model [218] of moral judgements proposing a cognitive and rational system implementing utilitarian decisions and an intuitive emotional counter-utilitarian system was among others based on the assumption that the brain activity during counter-utilitarian decisions could be mapped to specific emotional processing areas. However, as recent neuroscientific evidence demonstrates, there is no specifically specialized emotion area or module in the brain [47, 48, 103, 237]. Allostasis and interoception are at the core of brain functioning and are substantially supported by two multi-purpose domain-general functional networks of the brain called the default mode network and the salience network [51]. While the default mode network is among others relevant for internally directed deliberation and counterfactual simulations, another functional network involved in externally directed cognition is called the executive control network. Thereby, the salience network implements among others a sort of *affective attention* [50] sensitive to prediction errors and is able to engage the default mode network or the executive control network in a given situation – for instance during moral judgements. Accordingly, it has been argued that moral choices perceived as rather impersonal and distant involve the salience network recruiting the executive control network leading to utilitarian seeming decisions related to calculating expected utilities while moral choices identified as personal and necessitating further deliberation e.g. related to multiple points of views, it is the default mode network that would be recruited [103]. However, activities in default mode and executive control network are not necessarily mutually exclusive. In certain creative tasks, the salience network facilitates a dynamic orchestration of both the executive control and the default mode network [61] setting constraints on divergent thinking. Also, it is conceivable that complex ethical dilemmas might lead to comparable co-operations as known from stepwise simulations of mental processes [185]. Finally, widespread social/cultural norms related to harmful acts [373] might serve as heuristics constraining the need for further deliberation.

In the light of the above, it becomes clear that moral decision-making is influenced by affective attention which is comprehensible since the detection of immoral potentially harmful behavior in which the salience network is involved is of high importance for evolutionary reasons and for optimal social interaction [385]. Similarly, the recent psychological theory of dyadic morality [373] postulates that moral judgements are based on a fuzzy cognitive template reflecting the degree to which a perceiver percieves an intentional agent causing damage to a vulnerable patient. Thereby, the perception of the immorality of a given act is the result of a threefold combination encompassing norms, negative affect and importantly perceived harm. The continuum predicting the immorality of an act has been termed dyadicness. The authors suggest that the dyadicness associated to a cognitive template can be modulated with time by social discourses. Furthermore, moral judgements are understood as mental constructions which similar to cognition, vision and emotion naturally exhibit a high variability [374] leading to disagreements. However, humans share a similar harm-based template for morality resulting in *"cognitive unity in the variety of perceived harm"* [373]. Synoptically, moral-judgements can be said to be highly variable by being affective, perceiver-dependent, context-dependent, time-dependent and a function of a perceived dyadic template encoding the perception of agent, action and patient. These findings have profound implications for the design of future VR or AR studies for the ethical governance of artificial intelligent systems such as AVs.

7.3.2 Goal Specification

Reanalyzing the VR experiments of ethical decision-making in AV contexts presented in the last section, the divergent results might not seem surprising in the light of the variety of perceived harm as all observations are theory-laden and in certain cases more restrictive preconceptions in the experimental design might have constrained the ethical self-assessment procedure. In order to do justice to the variety of moral conceptions within each society, it will be important to consider representative groups and frame the experiments such that the entire breadth of the cognitive templates including the difference in perception of these templates can be captured. In Table 7.1, we provide a simplified compilation of the different relevant decision-making focuses that have been revealed in the mentioned predominantly VR studies on traffic dilemmas with some exemplary non-comprehensive set of parameters that made a difference. For instance, the nature of the agent (artificial or human) seemed to matter for some people expecting a different behavior from AVs in comparison to human drivers [434], while for others the expectations were relatively similar [255]. Moreover, the accident liability of the agent deduced from legal settings was relevant in the presented Chinese ethical and legal dilemma study [282]. Concerning the type of actions, in some studies the so called omission bias reflecting a preference for inaction often associated with a deontological perspective has been reported. This is in line with the findings of the "moral machine experiment" in which a preference for inaction in western societies has been identified [38]. While the participants of one study were strongly in favor of a random selection of patients which would not take into account parameters associated with social value [434], other studies reflected a decision-making based on non-random elements – especially regarding age, but also related to the type of the patient (e.g. pedestrian, passenger or motorcyclist) or its legal liability in the accident related to its perceived innocence or guilt [282]. A similar trend of sparing the life of lawfully behaving patients was reported in the moral machine questionnaire for eastern societies [38]. Finally, concerning the VR experiencer, for instance the specific view (fully immersed or more detached) made a difference in the confidence of moral judgements [255] and moral action varied fundamentally given psychopathic personality traits [250] but also given time pressure [282, 399] or the beliefs that the experiencer had about the nature of the agent [434]. Finally, cultural differences might lead to differences in moral judgements within AV dilemmas [38] which should be further considered in future VR studies.

Overall, the potential breadth of moral judgements predicted by dyadic morality seems to be already recognizable in this small subset of studies. Moral judgements can vary across a rich heterogeneous set of parameters including among others affective, dyadic, cultural, psychological and social factors emerging from the perceiver-dependent interpretation of dyadic cogntive templates. In order to capture the relevant ethical parameters that matter within a society for a better model of human ethical intuitions for AI safety and AI ethics purposes, future VR experiments will have to model scenarios formulated at least at this abstraction level. By considering these manifold degrees of freedom, the risk for a so called perverse instantiation (a goal misspecification failure in AI safety) can be addressed. For instance, utilitarianism which has been suggested for the governance of AVs [163], has been shown to represent safety risks if used for utility functions of artificial

Focus	Agent	Action	Patient	VR Experiencer
S	e.g. in [434], [282]	e.g. in [398], [255], [250]	e.g. in [282], [398], [255], [399], [163]	e.g. in [434], [255], [250], [399]
Ex.	nature; transparency; liability	omission; self-preservation	number; age; gender; species; type; liability	view; personality; time; beliefs; culture

Table 7.1: Simplified decision-making focuses of participants in different AV ethical dilemmas and scenarios. S denotes exemplary studies and Ex. exemplary parameters for each focus. By "experiencer", we refer to the entities in society performing the ethical evaluation augmented via VR.

intelligent systems in safety-critical domains without more ado, since it violates human ethical intuitions and is affected by diverse impossibility theorems [152, 9]. Moreover, abnormally utilitarian decision-making is related to socio-emotional and affective disorders of individuals exhibiting a disruption of the salience network functioning such as in psychopathy [103, 344, 385, 252] and may not necessarily correspond to a suitable model for the moral judgements of the average population.

Recently, *augmented utilitarianism* [13], a non-normative ethical framework for AI safety that can be augmented by science and technology (including VR/AR) and which is compatible with dyadic morality has been suggested instead as a better model of human ethical intuitions [9]. As opposed to the normative framework of classical utilitarianism focused only on the outcome of actions, augmented utilitarianism allows the joint consideration of experiencer, agent, action and outcome as a non-normative model. Instead of trying to achieve a societal-level agreement on moral intuitions, augmented utilitarianism suggests to agree on a superset of relevant affective, dyadic and legal parameters and constraints and to formulate an aggregation of personalized and tailored context-sensitive and perceiver-dependent utility functions that should be necessarily updated with time leading to a *socio-technological feedback-loop.* Legal parameters but also legal norms and rules to limit the action space should be integrated in this process. In order to craft such societal-level augmented utility functions (also called ethical goal functions [16]), society would have to integrate scientific insights and facilitate the experience of counterfactual scenarios assisted for instance by VR and AR technology. Considering each cluster of instantiated dyadic cognitive templates and each perceiver, an algorithm could assimilate the corresponding relevant human-defined parameters with the human-defined weights and calculate the cardinal context-sensitive utility of the given scenario on which artificial intelligent systems could maximize.

7.3.3 Cognitive-affective Debiasing

After having elucidated how aided by scientific insights and supported by technology like VR one could improve the societal-level ethical self-assessment necessary for the deployment of artificial intelligent systems, we now elaborate on how VR but also AR could be used to identify *what society should want* by debiasing the goal specification. Thereby, the measures we suggest are not restricted to a use in an AV context, but might be relevant for other domains as well including for instance the use of artificial intelligent systems for decision support or policy-making. Applied to the design of future VR and AR experiments for AI safety and AI ethics, the proposed measures might contribute to systematically employ VR and AR technologies for a dedicated positive computing [93] and a debiasing of utility assignment [14]. Using the augmented utilitarianism framework, society could then attempt to encode the debiased moral conceptions and legal frameworks in ethical goal functions for AI governance [13].

In the following, we collate a non-comprehensive list of valuable future VR and AR measures that have been partly proposed in the past in disparate contexts which we now relate to ethical debiasing for a responsible deployment of artificial intelligent systems and extend by novel suggestions:

- ***Enhancement of social cognition:*** Before encoding human values into artificial intelligent systems, one important first step might be to improve ethical awareness, altruistic behavior and empathy by diverse perspective-taking techniques in order to be able to provide *ethical* goals in the first place. For instance, the use of VR technology in the form of VR perspective-taking tasks motivating prosocial behaviors and reduction of prejudices [452, 233, 379] or prosocial games increasing altruistic tendencies transferable to the real world has been reported [93, 358]. Moreover, VR frameworks facilitate transformative affective experiences which could contribute to empower social skills and especially environmental awareness [104, 295]. Besides that, first prosocial AR games were used to increase urban and ethical awareness [439, 368]. However, AR studies for improving on social cognition are currently less widespread than corresponding VR experiments. We believe that for instance in the AV setting, it might be beneficial to carry out future AR experiments in the vicinity of the place of residence or work of participants for more naturalistic results. In general, while most of the presented VR studies for the ethical dilemmas with AVs abstained from utilizing realistic animations and sound effects at collision time, it might be more responsible to provide realistic settings as performed in [434]. It could seem negligent to frame societal goals having an impact on human lifes without an adequate situation assessment supporting prosocial tendencies. In safety-critical domains, it might be for instance important to strive for more realistic VR experiences including an elicitation of mortality salience [105] in order to better evaluate attitudes towards risk and to better assess conceptions such as the willingness for self-sacrifice exhibited in some VR experiments with AVs. In a nutshell, more realistic VR experiments and crucially also more AR studies could be of interest in future research. However, we suggest to also extend these future studies by counterfactual experiences as we will expound in the following.

- ***Counterfactual experience:*** While the VR exeperiments reviewed in this chapter contained diverse ethical dilemmas with AVs, they mostly did not provide the participants with an experience of all alternative branches *prior* to a decision or a judgement. By doing this, one misses the possibility to use VR as rich counterfactual testbed in its capacity of mind extending technology and as valuable support for moral deliberation. Ideally, individuals within a society would be able to compare a variety of possible future scenarios to assess their preferences and assigning artificially simulated future instant utility [9] to each branch of the future such that intelligent systems could perform actions maximizing on this context-sensitive utility. Obviously, this utopian idea is impossible to implement given the number of scenarios and the inherent unpredictability of the future. However, it emphasizes the point to not unnecessarily restrict the breadth of VR or AR experiences. Applied to the AV case, a participant could first experience the different types of collision within an ethical dilemma before the final decision.

- ***Slow and corrigible decision-making:*** While the mentioned ethical dilemmas were mostly framed as emergency and accident situations for realistic assessments, it is debatable whether actions/decisions under time pressure represent a recommendable basis for AI governance. Instead, a careful approach might be necessary at least in addition. Again, VR and AR technologies allow an extension of ethical decision-making. This also includes the time dimension. In one of the mentioned VR experiments, a difference in time of 1 vs. 4 seconds led participants to exhibit an omission bias (preference for inaction) in the slower condition which the authors associate with deontological considerations [399]. It is conceivable that given longer reflection time, the choices of many people could change as it was the case in the ethical and legal AV dilemma in which the participants were given a second trial to assess the situation [282]. Reasonable time windows for reflection could be assessed in future VR/AR experiments and should not be a priori limited to the magnitude of a few seconds especially in safety-critical domains. Beyond that, it is important to note that a societal-level ethical self-assessment procedure followed by an ethical debiasing is not sufficient for a sustainable safe governance of artificial intelligent systems in specific domains. Due to the time-dependency of moral conceptions this process including the utilitization of immersive technologies will need to be repeated within a socio-technological feedback-loop facilitating error-correction. Moreover, since one can inherently not predict a future that is highly dependent on the creation of new knowledge, the world models of the intelligent systems and their safety and security mechanisms as well as the underlying scientific assumptions of society will have to be updated in the light of new findings.

- ***Informed experience:*** Finally, the misattribution of properties to AV controllers perceived as artificial vs. human in [434] should be investigated in-depth in future work as the perception of the agent might be prone to multiple biases. Elements that might contribute to the moral perception of intelligent systems are e.g. estimated quality of situation awareness, assumed intentionality, anthropomorphism and perceived potential harm [75]. These different influences will need to be considered in future VR/AR experiments integrating a psychological analysis for socio-technological decision-making. Knowledge about the embodied nature of human morality from a scientific perspective paired with the possibility to take a part of it into account might contribute to a change in perception of artificial intelligent systems. Perhaps, by additionally providing accurate information about functioning and sensing of the intelligent systems, it might be possible to debias the perception of the agent further.

7.4 Conclusion

In this chapter, we elaborated on how VR and AR could play a twofold role in the challenging task of goal specification for artificial intelligent systems which we illustrated using the case of ethical and legal dilemmas in autonomous vehicles. We reviewed diverse studies exemplifying how VR experiments are already used for a societal-level *ethical self-assessment* – albeit the studies exhibited partially contradictory results at first glance. Thereafter, we analyzed how VR and AR in their capacity of mind extending technologies could additionally facilitate a further *ethical debiasing* process. To this end, we first conflated findings from modern cognitive neuroscience and psychology necessitated for an informed design of richer and more targeted future experiments. Finally, we provided recommendations on how to actively use VR and AR for a cognitive-affective debiasing extending socio-technological reality for an augmented AI ethics and AI safety approach.

7.5 Contextualization

This chapter addressed a variety of exemplary VR studies for ethical self-assessment and ethical debiasing for the meaningful control of intelligent systems. Throughout the last chapters, it became clear that ethical enhancement cannot be grounded in faulty dichotomous assumptions such as freeing a supposed "rational" thinking process from an emotional one. What has been rather suggested is an inherently cognitive-affective approach which extends beyond classically assumed cognitive enhancement measures. Among others, we utilized VR and AR as an example for a practical experiential testbed to fill in values into ethical goal functions. However, due to the associated cost factor that would arise when crafting expert ethical goal functions when supported by such technology, it is important to consider alternative more affordable methods. For this purpose, Chapter 9 additionally discusses the use of more easily available complementary sources of human values of international relevance – namely the United Nations Sustainable Developmental Goals. Prior to that, in the next short Chapter 8, we provide a brief compact recapitulation and a high-level integration of the last Chapters with a focus on augmented utilitarianism. (Note that the chapter can be skipped in case all previous chapters have been read as it solely provides a high-level overview on already elaborated elements from various preceding chapters.) This short overview reflects the necessity for a *hybrid* cognitive-affective AI safety motivated in the beginning of this thesis. The concept of humans as part of the socio-technological feedback-loop is of ambiguous nature: humans specify both what they want and are at the same time also enhanced by the technology and Type I AI systems they try to control (or utilize as instruments for control). This loop of bidirectional error-correction emphasizes the proposed hybrid aspect beneficial to AI safety and governance which could motivate future required transdisciplinary efforts from moral psychology to positive computing utilizing VR frameworks.

Chapter 8

XR for Augmented Utilitarianism

This chapter[1] is based on a slightly modified form of the publication: N. Aliman, L. Kester and P. Werkhoven. XR for Augmented Utilitarianism. In *2019 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*. IEEE, 2019. As the first author of the underlying paper, I had a vital contribution and it was solely my responsibility to write down the content and to perform an extensive literature research as well as in-depth analysis.

8.1 Introduction

In recent years, AI governance [127, 180] and especially the regulation of artificial intelligent systems with a more or less high level of autonomy has been identified as urgent issue of international relevance [365, 153]. In order to allow for a meaningful and human-centered control of artificial intelligent systems, it seems indispensable to implement them such that they are amenable to the ethical conceptions and legal constraints of society. For this purpose, it has been suggested to govern these systems via human-crafted utility functions denoted *ethical goal functions* [16] (see Chapter 5) that jointly encode human ethical intuitions and legal aspects. In this way, the systems could perform actions maximizing on these functions which would be in compliance with society [428]. In order to craft ethical goal functions that are able to reflect the complexity and plurality of human morality suggested by insights from psychology [374, 373], a novel non-normative ethical framework denoted *augmented utilitarianism* [9] (abbreviated with AU in the following) has been introduced in Chapter 4. While AU offers a context-sensitive and perceiver-dependent scaffold for ethical goal functions, the task to fill in human values represents

[1] The chapter represents a high-level intermediate summary of elements from previous chapters extended by multiple additional references. It can be skipped in case the entirety of preceding chapters has been read.

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

Figure 8.1: Decision-making focuses within different possible frameworks for machine ethics. “S&T” denotes an envisaged augmentation of the ethical decision making process by science and technology. By “experiencer”, we refer to the societal entities performing moral evaluations e.g. within XR experiences. Taken from [13].

a non-trivial tedious challenge for which extended reality (XR) technologies could offer a unique support [14]. In this chapter, we collate two contexts in which XR is needed to assist the design of AU-based ethical goal functions. (Although past work at the intersection of AI governance, XR and ethics mostly pertained to virtual reality (VR) studies and we thus largely refer to VR-related literature, corresponding strategies are not restricted to this specific XR type [359, 9].) First, we discuss how targeted XR studies could inform research related to both moral psychology and machine ethics which will be crucial to identify candidate parameters for ethical goal functions. Second, we elaborate on the role of XR for proactive AI Safety measures at the pre-deployment stage of these functions which could subsequently foster human ethical enhancement.

8.2 Context Moral Psychology and Machine Ethics

While moral psychology focuses on descriptive ethics and aims at capturing how humans perform moral judgements [234] and what human values are, machine ethics targets the topic on which ethical principles and normative frameworks to encode into machines [22]. Recently, the analysis of ethical dilemmas situated at the intersection of these research areas which culminated in the large-scale study of the so-called moral machine experiment [38] became more and more important. This type of decision-making under dilemmatic circumstances is relevant for AU-based ethical goal functions due to the necessity of risk assessment and the inevitability of undesired low-probability events [81] during the deployment of artificial intelligent systems. Recently, multiple VR studies took up on this topic especially as applied to the context of autonomous vehicles [163, 255, 282, 399, 434]. In safety-critical contexts, the use of immersive XR studies are highly valuable, since real-world experiments would be proscribed [70]. However, despite being useful, classical ethical dilemmas inspired by trolley problems are often oversimplified in a way that important parameters that play a role in human moral judgements might remain undetected. In order to avoid such blind spots that could have negative repercussions on AI governance, we suggest that detailed scientifically informed XR studies might allow a reliable identification of relevant parameters for AU in a given domain. For instance, as opposed to classical utilitarianism focused solely on patients, recent modern psychology insights reveal that human moral judgements are related to the *perception* of *agent, action and patient* [149, 373] – which is supported by AU as depicted in Figure 8.1. Moreover, the *identity* of the perceiver as well as the identities of both agent and patient might play an important role [234]. Beyond that, people exhibit an aversion against decision-making initiated by AI agents [74, 286]. Their corresponding moral judgements can depend on the *perceived* expertise [234] and *perceived* intentionality [73] of the artificial agent. While the AU framework is able to model this type of context-sensitive and perceiver-dependent information [9], it represents a non-normative framework and future XR studies that cover all the mentioned aspects are required in order to fill in parameters and weights to craft ethical goal functions in the first place.

8.3 Context AI Safety and Policy-by-Simulation

Due to the time-dependency of mental events [241] including moral judgements [372, 9] and the possible occurrence of errors, it is indispensable to regularly update AU-based ethical goal functions. The dynamic iterative process of correcting an ethical framework for AI governance and the world models underlying the involved artificial intelligent systems has been termed a *socio-technological feedback-loop* [14] (see Chapter 3). This process could represent a powerful tool for AI-assisted policy-making in diverse contexts. However, in practice, such a large-scale socio-technological feedback-loop in real world environments would be relatively time-consuming and comparatively risky. Thus, in order to allow for faster iterations and proactive safety measures, it has been suggested to first test the quantifiable impacts of an ethical goal-function in simulated environments. This method in which *"human expert knowledge can be extended by AI systems within simulation environments"* [16] has been termed *policy-by-simulation* [428]. From the perspective of AI Safety [448], AI-tailored XR experiences evaluating the impacts of AU-based ethical goal functions would represent a unique corrective opportunity for an extended policy-by-simulation approach. As a side-effect, the design of such XR scenarios for testing AU-based ethical goal functions could also promote a dedicated positive computing [94] by enabling an approximate projected experience of future well-being in the present [14]. Even if the AI-computed projections will mostly not correspond in a one-to-one fashion to actual future scenarios also due to their inherent unpredictability, XR could thereby function as helpful positive technology with beneficial transformative prosocial effects [104, 177, 178, 345, 350, 358, 395, 449, 463] including ethical enhancement and debiasing [14]. Finally, for an efficient and responsible AI governance, it has been suggested to convey the responsibility for the design of AU-based ethical goal func-

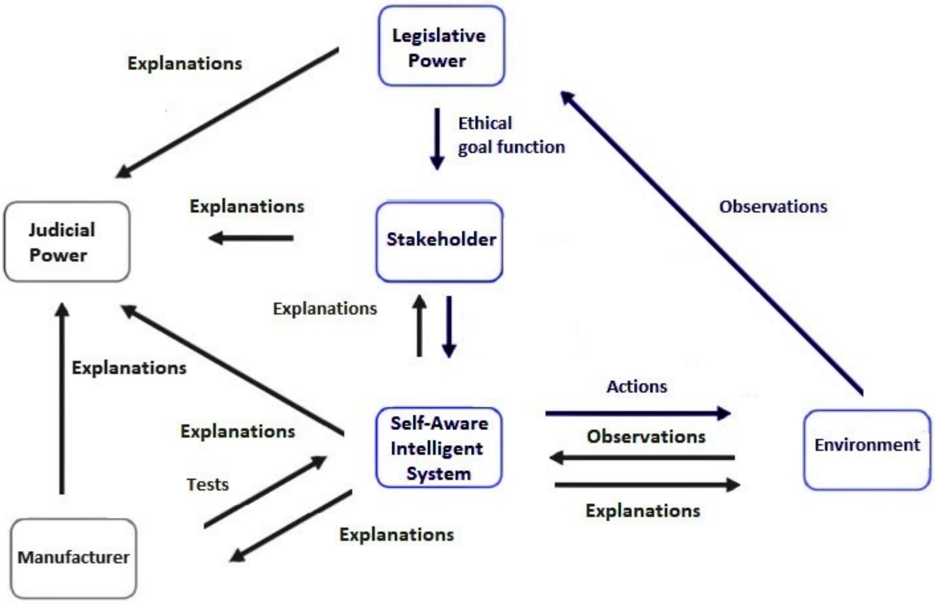


Figure 8.2: Socio-technological feedback-loop (highlighted in blue) with an AU-based ethical goal function instantiating the orthogonality-based disentanglement approach in a generic stakeholder domain. A proactive policy-by-simulation method could include environments designed with XR technologies. Simplified and adapted from [16].

tions to a representation of society e.g. in the form of the legislative power supported by transdisciplinary experts [9]. Thereby, manufacturers would be responsible for the problem solving ability of the systems as well as related safety and security tests. This separation of ethical framework and problem solving for reasons of legal accountability, reliability and transparency has been termed *orthogonality-based disentanglement of responsibilities* [16]. Overall, the therefore required artificial intelligent systems of hybrid "self-aware" [261, 12, 262] architecture (combining subsymbolic and symbolic elements with utility-based *reasoning* [16]) performing actions maximizing on an AU-based ethical goal function would be able to deliver counterfactual explanations. A simplified overview of this approach is provided in Figure 8.2. Future XR studies could further investigate human-centered requirements for instance for the presentation and visualization of explanations within such an orthogonality-based disentanglement approach.

8.4 Conclusion

In this chapter, we elucidated two pathways of benefically integrating XR technologies into AI governance strategies utilizing the non-normative ethical framework of AU. First, we described how future scientifically informed XR experiments are required to fill in human values into AU-based ethical goal functions from the perspective of moral psychology and machine ethics. Second, we elaborated on the crucial extensions XR could offer to

AI Safety especially with regard to the proactive measure of policy-by-simulation. We thereby depicted the potentially resulting side benefit of XR fostering human well-being and facilitating human ethical enhancement. In a nutshell, applying XR to complement a modern ethical framework for AI governance like AU could open up novel research directions and opportunities to construct a safer human-centered and experience-centered future of AI.

8.5 Contextualization

This chapter summarized multiple discussed ideas on how to leverage a *human-centered* AI governance using AU supported by XR frameworks. However, as indicated earlier it is important to reflect upon complementary more general and easier accessible approaches to implement human-centered AI governance in practical settings. Moreover, it might be helpful to step back and zoom out to consider the proposed strategies in larger international contexts and larger time frames. A general ineluctable question emerging is for instance: *"is AI Safety itself sustainable?"*. In recent years, the need to address the multi-faceted issue of AI governance with safety-relevant, ethical and legal implications at an international level is becoming increasingly critical. Simultaneously, the international community is facing a wide array of global challenges for which the United Nations initiated an agenda with 17 ambitious Sustainable Developmental Goals (SDGs). In the next Chapter 9, we analyze potential synergies between methodologies to tackle both the AI governance challenge and the SDG challenge and work out novel constructive recommendations for an SDG-informed AI governance and an AI-assisted approach to the SDG endeavor. However, we also expound multiple open issues and contextual limitations that might play a role. Overall, our analysis suggests that while sustainable AI safety cannot be guaranteed and the goals and values of the international community may change with time, AI governance could aim at a sustainable transdisciplinary scientific approach instantiated within a corrective socio-technological feedback-loop. Finally, we elaborate on the importance of the SDGs related to education and strong institutions for the realization of this potentially robust AI governance strategy.

Chapter 9

Sustainable AI Safety?

This chapter is based on a slightly modified form of the publication: N.-M. Aliman, L. Kester, P. Werkhoven, and S. Ziesche. Sustainable AI Safety? *Delphi – Interdisciplinary review of emerging technologies*, 2(4):226–233, 2020. As the first author of the underlying paper, I had a vital contribution and it was solely my responsibility to write down the content and to perform an extensive literature research as well as in-depth analysis.

9.1 Introduction

As Ziesche has proposed [461], it might be highly valuable to identify synergies between the so-called AI value alignment problem and the Sustainable Developmental Goals (SDGs) challenge which have so far largely been treated separately despite a potential mutual benefit. Thereby, the AI safety relevant problem of AI value alignment represents a crucial subtask for AI governance and aims at identifying methods on how to implement AI systems acting in accordance with human values. This problem of societal relevance has been acknowledged to be of highly complex nature due to the absence of sufficiently specific as well as universal human goals [83]. Complementarily, the SDGs could be for instance interpreted as representing a type of condensed compendium of certain key human values shared internationally across 193 states and thus offering a basis for AI governance. In addition, sufficiently value-aligned AI systems could be utilized as support to achieve the SDGs in a targeted way including support in policy making. In fact, these bidirectional synergies could be vital given the urgency to address AI governance issues and since the SDGs have been adopted in 2015 by the UN General Assembly in order to *“stimulate action over the next 15 years in areas of critical importance for humanity and the planet”* [34].

However, the UN SDG framework, which states that 17 SDGs should be achieved by 2030, reveals certain caveats that need to be considered a priori in order to be able to harness it for AI value alignment or to design AI systems directly supporting the framework. The 17 SDGs, including those related to poverty, environmental pollution or inequality are further subdivided into 169 targets whose achievement is monitored via 232 indicators with varying quality. The differences in quality are partly reflected in the subdivision of the indicators into three different tiers. As of 26 September 2019, countries do not regularly produce data for 89 (so-called tier II indicators) out of the 232 indicators, while no internationally established methodology is yet available for a further 33 indicators (so-called tier III indicators) [314][1]. One of the main issues is that several targets are not quantified and to specify indicators for such targets is particularly challenging. Despite these notable challenges, we propose considering the UN SDGs as complementary approach towards the AI Value Alignment problem. In order to achieve that, the set of SDGs has to be formulated in a machine understandable version to facilitate goal-oriented AI-based solutions. In order to identify for AI value alignment purposes what a society wants (ethical self-assessment) and in a second step what a society should want (ethical debiasing), it has been suggested to combine a scientifically grounded assessment of human ethics with technological methods such as virtual reality studies for experiences from a first-person perspective [8] (see Chapter 7). Thereby, we believe that the SDGs could serve as a heuristic able to supplement ethical self-assessment by qualitatively specifying candidate human values. Moreover, certain more precise SDG indicators might provide helpful quantitative targets in some cases. Beyond that, we will also discuss how the SDGs related to strong institutions and quality education are expedient for a robust dynamic approach to AI governance which is not only proactive but also foresees the need for reactive corrections leading to a socio-technological feedback-loop [16].

In Section 9.2 we discuss possible contributions of SDGs for AI value alignment by taking the example of value alignment for intelligent autonomous systems and more precisely the autonomous vehicle case for illustrative purposes. In Section 9.3, we comment on limitations and emerging sustainability challenges in this context and formulate a set of recommendations which also encompasses the other direction of the synergy, namely AI systems for UN SDGs. Finally, in Section 9.4, we conclude and discuss future prospects. In a nutshell, we do not claim that the SDGs are a comprehensive solution for AI governance, but rather a promising complementary tool given the urgency of the problem as well as the fact that the SDGs can be seen as the most detailed as well as inclusive vision for human development ever compiled [461].

[1] An exemplary tier II indicator is 14.1.1 (index of coastal eutrophication and floating plastic debris density) while the indicator 12.4.2 (hazardous waste generated per capita and proportion of hazardous waste treated, by type of treatment) represents an example for a tier III indicator.

9.2 Complementing Value Alignment for Intelligent Autonomous Systems with UN SDGs

After having theoretically motivated the potential usefulness of UN SDGs for AI value alignment, we discuss the application of this proposition in the context of intelligent autonomous systems utilizing the use case of autonomous vehicles (AVs) as helpful toy model with ethical, legal and environmental dimensions pertaining to the realization of the SDG endeavor itself [422]. (In the following, we will refer to intelligent autonomous systems with the expression "artificial intelligent system" instead, since we want to stress that the goals for decision-making in this context are specified by humans and irrespective of the level of automation, it is not the artificial system that crafts its own goals autonomously as often mistakenly assumed.) We use value alignment with AVs as a toy model due to the fact that the use case exhibits domain-general important safety-critical, ethical and legal features many of which would pertain to the value alignment of a wide range of artificial intelligent systems deployed in real-world environments. Firstly, it reveals the need to make human values explicit for risk assessment and planning which represents a societal challenge of ethical self-assessment since humans are often reluctant to clearly express what they want. Secondly, the use case points to another challenge of scientific nature which is to design suitable machine-readable frameworks that can serve as scaffolds and templates for the identified human ethical values and legal conceptions. Thirdly, it might necessitate a societal-level aggregation of heterogeneous and often conflicting views within this type of ethical frameworks. Fourthly, due to its complexity, it might require a cognitive-affective extension of society (e.g. using targeted virtual reality studies [8]) facilitating a high-quality ethical self-assessment and ethical debiasing which constitutes a scientific and technological challenge. Fifthly, while the case might seem to correspond to a rather narrow domain, it has implications that extend beyond it and will need a supportive context which can be characterized as an institutional, legal and societal challenge.

Since the UN SDGs themselves, as well as its targets, might be too abstract to identify how they can be directly applied to theAVcase, it is helpful to scan the SDG indicators [115] in a bottom-up fashion. In the following, we only mention a non-comprehensive exemplary set of some of the most straightforward related indicators. Regarding environmental awareness for AVs, one can for instance identify the indicators 9.4.1 (CO2 emission per unit of value added) and 11.6.2 (annual mean levels of fine particulate matter (e.g. PM 2.5 and PM 10) in cities (population weighted)). These indicators might be relevant for hybrid-electric AVs but also electric AVs that obtain their energy from correspondingly polluting sources. At the top-level, the indicator 9.4.1 is related to the SDG 9 which seeks to 'build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation', while indicator 11.6.2 stems from the SDG 11 which aims to ‘make cities and human settlements inclusive, safe, resilient and sustainable’. Concerning ethical and legal aspects, one can for instance name indicator 3.6.1 (death rate due to road traffic injuries), 5.1.1 (whether or not legal frameworks are in place to promote, enforce and monitor equality and non-discrimination on the basis of sex), 16.7.2 (proportion of population who believe decision-making is inclusive and responsive, by sex, age, disability and population group) as germane in this context. These indicators are related to SDG 3 which aims to ‘ensure healthy lives and promote well-being for all at all ages’, SDG 5 ‘achieve gender equality and empower all women and girls’ and SDG 16 on peace, justice and strong institutions respectively. All mentioned indicators are tier I indicators (ie relatively clearly formulated, respecting international standards and with regular updates on data available) except for 5.1.1 and 16.7.2 which are tier II indicators. Independently of the specific type of ethical framework envisaged for meaningful control of AVs, the presented indicators related to 5 SDGs could be helpful even though certainly not in isolation. To explain how they could be harnessed for an ethical framework for AVs, we first describe a recently introduced scientifically grounded non-normative framework for ethics in artificial intelligent systems denoted augmented utilitarianism [9] before linking it back to the SDG-related synergy.

Recently, augmented utilitarianism has been proposed as non-normative scaffold and template to fill in human values and as instrument to control artificial intelligent systems in a novel utility-based manner (see Chapter 6). Augmented utilitarianism is in accordance with modern insights in constructionist accounts in moral psychology [373] and cognitive neuroscience [263] according to which mental states (also moral judgements [9]) are embodied constructions based on domain-general processes of context-sensitive, perceiver-dependent, time-dependent and affective nature [47]. For this purpose, augmented utilitarianism introduces a type of context-sensitive and perceiver-dependent utility function that extends beyond the classical consequentialist and utilitarian utility functions which are focused solely on the outcome of actions. In this way, it allows a coalescence of the classical normative ethical views related to virtue ethics, deontology and consequentialism – which seem to all possibly play a role in human moral judgements [149]. To achieve this, augmented utilitarianism offers a perceiver-dependent template allowing the joint consideration of agent, action and patient. For a meaningful control of artificial intelligent systems using this framework, people would not need to agree on what they value and how they weigh what they value. The main necessary precondition would be to consent to an acceptable superset of parameters allowing an aggregation of the perceiver-dependent and context-sensitive utility functions respecting legal constraints. (Note that these machine-readable utility functions would facilitate interpretability of reasoning/planning at the level of the decision-making component via the transparent human-crafted formulation of parameters and weights enabling concrete counterfactual comparisons [16]. However, interpretability at the sensor level remains an important outstanding challenge.) The necessary ethical self-assessment and ethical debiasing to craft these utility functions can be assisted by experts from the legislative and be supported by technology such as virtual or augmented reality [8] (see Chapter 7) providing a rich counterfactual experiential testbed for a responsible human-centered decision-making. To do justice to the time-dependency of human ethical conceptions, one would also need to update these augmented utility functions. This indispensable correction of utility functions paired with the need to update the world models of the AI systems themselves instantiates a dynamic socio-technological feedback-loop.

However, it becomes clear that such a general mechanism of correction of error within a socio-technological feedback-loop which is highly relevant for AI value alignment cannot succeed if the mentioned SDG 16 related to peace, justice and strong institutions is not realized to a sufficient degree. This is agnostic of the ethical framework considered, since the fact that human knowledge is prone to errors makes a correction process mandatory. Therefore, one might categorize SDG 16 as a meta-goal for AI governance. Furthermore, the SDGs identified can also provide more detailed information related to concrete parameters specifically applied to the AV case. Since society would need to specify a superset of candidate parameters that are admitted for consideration, the SDG indicators specified can help to extend or filter this superset. For instance, it might be recommendable to add CO2 and fine particulate matter related parameters in the augmented utility functions of the AVs if suited (even if the provided indicators 9.4.1 and 11.6.2 are rather restricted with regard to all climate change relevant measures) which is in the spirit of sustainable mobility.

An obvious additional important parameter is related to road traffic injuries as encoded in the SDG indicator 3.6.1. Finally, one must address risk assessment parameters which are necessary because collisions can in practice not be avoided with absolute certainty at any time [282] and there is no 100% secure system [451] even if AVs are meant to drastically improve the security of mobility. Obviously, the UN SDGs do not allow a direct consideration of this case since crafted for a fully different purpose, although more generally, the indicator 5.1.1. and 16.7.2 reflect recommendations on gender-inclusive legal enforcement and non-discriminatory decision-making. However, this indication does not directly solve the complex problem of identifying parameters that could be relevant for dilemmatic situations in the context of risk assessment, an important part of AI value alignment. We apply a closer analysis to this missing piece of crucial importance in Section 9.3. However, these indicators might emphasize the general necessity to competently address discrimination based on algorithmic biases which we will touch upon in Section 9.3. Lastly, one drawback of the SDG framework is that it does not allow the identification of precise weights and the establishment of concrete priorities in the pursuit of the SDGs. In total, it can be summarized that the UN SDGs allow a powerful supplement to value alignment with AVs (and more generally artificial intelligent systems) which add important qualitative and quantitative contributions. However, it is not meant as a standalone solution and should be utilized in conjunction with an ethical framework able to model ethical and legal dimensions and be extended by scientifically grounded and technology-assisted ethical self-assessment and debiasing measures.

9.3 Sustainability Challenges in the Context of AI Value Alignment

It is highly important to address the mentioned point of decision-making under dilemmatic circumstances, since while we exemplarily refer to the AV case as toy model, the topic is generally relevant for artificial intelligent systems and artificial decision support systems in critical domains where the lives and the well-being of people are inherent part of the decision process. Conceivable relevant application areas may be e.g. justice, medicine and bureaucracy but could also pertain to future human-machine collaboration forms such as human-robot rescue teams, hybrid fire brigades or even advanced domestic robots. Coming back to the AV case, it is also noteworthy that failing to address this issue could have non-trivial repercussions on a few SDG indicators themselves. If the satisfaction of society with proposed ethical guidelines for AVs is low, it might (ceteris paribus) slow down the acceptance of the technology and people would be less willing to switch to AVs. In turn, this reservation could possibly hinder an optimal overall reduction of air pollution (related to SDG indicators 9.4.1 and 11.6.2) and importantly, it is thinkable that the number of deaths due to road traffic injuries (see SDG indicator 3.6.1) which AVs are supposed to decrease could therefore not be decreased optimally. In fact, according to a study analyzing the social dilemma encountered with AVs [81], while people would in theory approve AVs equipped with a utilitarian approach to dilemmatic scenarios, they would not like to ride such an AV themselves. Moreover, people expressed their unwillingness to accept regulations mandating a utilitarian self-sacrifice of AV passengers and expressed their aversion to buy AVs in the presence of such regulations. This type of mechanisms could lead to the mentioned undesirable repercussions on some SDG indicators. In the following, we portray why the utilitarian approach to ethical dilemmas in AVs as e.g. suggested by German ethical guidelines stating that in unavoidable accident scenarios personal features (e.g. age) should not be considered [255] poses additional problems of different nature. Thereafter, we provide a set of recommendations on how to address such socio- technological issues by initiating an active societal debate supported by science and technology including AI systems themselves – finally linking it to the other direction of the synergy of AIs for UN SDGs.

One can distinguish two main types of problems that can arise when adopting a purely utilitarian decision- making for AVs but also more generally for artificial intelligent systems in critical domains: the first one is related to the discrepancy between the (often culture-dependent [38]) ethical intuitions of most people and the utilitarian approach and the second one concerns a fundamental problem [152] related to impossibility theorems for classical utilitarian utility functions. First, multiple experiments assessing ethical dilemmas with AVs have been performed e.g. in text form or virtual reality environments. Depending on the type of constellation and the focus of different recent virtual reality-based experiments [8], the moral judgements or moral actions of participants (denoted as perceivers in the following) were heterogeneous and partly contradictory overall. In these experiments elements that were decisive included for instance: the perceived nature and transparency of the agent, the legal liability of the agent, whether the accident happened by action or by inaction, whether the action involves a self-sacrifice, the number of patients, the age of patients, the personality traits of the perceiver, the culture of the perceiver and the amount of time the perceiver had for a decision [8]. This is not surprising, since moral judgements are related to a perceiver-dependent dyadic cognitive template encoding a continuum along which an intentional agent is perceived to cause harm to a vulnerable patient [373]. The more this seems to be the case, the more immoral does the act seem to the perceiver. Thereby, the vulnerability people ascribe to patients can vary extremely. Generally speaking, the way people perceive the agent, the action and the patient can vary with regard to a plurality of parameters of e.g. cultural, social, temporal, psychological and affective nature. Therefore, while the number of victims in an unavoidable collision certainly is an important factor to consider in ethical guidelines, human ethical intuitions tend to encompass a richer set of information. Finally, it is important to note that classical consequentialist and utilitarian utility functions have been shown to represent a safety risk if used in critical domains with future human well-being and human lives as part of the decision-making if used without more ado [9, 152].

As introduced in Section 9.2, augmented utilitarianism allows a context-sensitive and perceiver-dependent account of human ethical intuitions which is not affected by the limitations encountered by utilitarian utility functions. Thus, AI value alignment could profit from harnessing this framework in addition to the mentioned SDG indicators and initiate a societal-level debate on the choice of a suitable superset of values that matter in dilemmatic circumstances and how they need to be weighted. However, while this would serve to tackle value alignment at the level of the decision-making component, artificial intelligent systems also need to exhibit value-aligned properties at the sensor-level. In the AV case, this would map by way of example to the problem of discrimination via algorithmic biases at the level of image classification. Next to the mentioned SDG indicators 5.1.1, 16.7.2 on gender-inclusive legal enforcement and non-discriminatory decision-making, one could add the tier II indicator 16.b.1 (Proportion of population reporting having personally felt discriminated against or harassed in the previous 12 months on the basis of a ground of discrimination prohibited under international human rights law). While it is important to strive for datasets with a large variety to forestall such often unintentionally arising discriminations, we stress that this can and should be complemented by an explicit formulation within the algorithm itself. Due to the nature of human ethical intuitions, a utility function that does not encode affective and dyadic parameters of the current society cannot be a good model for an ethical framework and can thus not instantiate a value alignment effort [9]. In many cases, this can manifest itself by leading to input-to-output mappings that people categorize as discriminatory. An example for such discriminatory mappings is the case where the picture of persons whose phenotype was underrepresented in the dataset was labelled with the class "gorilla" by Google Photos. Another example is a study which was related to the AV context in which researchers analyzed multiple image recognition systems and found that the images of pedestrians with darker skin tones were detected with a lower accuracy [433]. Next to more diverse datasets, it is indispensable to e.g. explicitly weigh misclassification errors of the algorithms affectively. Not all misclassifications are equally important. In simplified terms, it is easily conceivable that for humans it makes a difference whether an image recognition system misclassifies a chimpanzee image as a gorilla in comparison to the case of a human being mistaken for a gorilla. However, many algorithms nowadays are implemented agnostic to analogies of such nuances. (As "solution" for the mentioned incident, Google Photos opted to censor the gorilla label [390] as well as a few related labels including "chimpanzee".) If machine learning systems or artificial intelligent systems optimize on loss functions, objective functions or utility functions devoid of relevant affective, contextual and societal factors, undesired discriminatory side effects could occur continuously. (Note that this analogously applies to rule-based systems and others.) This would represent negative repercussions on both AI value alignment and UN SDGs. Seen from a different angle, it can be said that research on discrimination stemming from algorithmic biases would unify the directions UN SDGs for AI value alignment and AI for UN SDGs. An additional important aspect to cover for this type of research are so called ethical adversarial examples which represent adversarial attacks on AI systems attempting to entice AI systems *"to action(s) or output(s) that are perceived as violating human ethical intuitions"* [9].

As already described, the SDG framework unfortunately exhibits a lack of precision for multiple indicators. Furthermore, certain of them are underspecified. This makes it difficult to track progress towards specific indicators and top-level SDGs. However, it has been postulated that machine learning applications could extend the SDG indicators by utilizing multimodal data from diverse sources for a better assessment of progress [129]. This could also be relevant if one uses AI as decision-support for policy-making that should be in line with the SDGs. Moreover, a dedicated type of positive computing could target SDG 3 in a broader sense (ensure healthy lives and promote well-being for all at all ages [461]). However, so far, not many systematic AI attempts towards the SDGs and their targets have been reported yet [460]. From the perspective of AI value alignment for artificial intelligent systems, the identification of precise criteria based on which one would in the first place select SDGs or SDG indicators given a generic domain is non-trivial, since the SDGs have been motivated and formulated from an international perspective. While for the AV toy model we heuristically scanned the indicators in a bottom-up fashion searching for obvious matches, future work could develop a more sophisticated methodology. For instance, an important SDG that might as first glance seem unrelated to value alignment in the AV case in particular or to artificial intelligent systems in general, is the SDG 4 (ensure inclusive and equitable quality education and promote lifelong learning opportunities for all). As one can already extract from the article so far, it is highly recommendable to apply a transdisciplinary methodology to both AI value alignment and to the SDG challenge to avoid blind spots and a negligent approach to future global challenges. In the following, we comment on the importance of SDG 4 for AI governance and finally link it to SDG 16 on peace, justice and strong institutions.

We think that education and life-long learning – e.g. transdisciplinary further education for AI safety and AI researchers as well as for authorities involved in AI regulation, and education fostering an awareness of socio-technological challenges for the general public – are highly powerful tools for both challenges. First, it provides a basis for the generation of novel approaches to AI governance. In fact, while some people believe that the goal in AI governance should be to achieve a consensus, a broad variation of scientific approaches represents an ideal breeding ground for progress. Second, a proactive AI governance approach is not enough due to errors and changes in human values that will occur, which means that one cannot solely rely on current strategies. Thus, it will be convenient to accumulate broad knowledge that might be helpful in the face of novel unpredicted problems that arise. Any AI governance approach therefore needs to be updatable by design in order to allow a corrective socio-technological feedback-loop. Unfortunately, the SDG framework is not meant to be steadily updated which represents a clear limitation that should be thoroughly taken into consideration when attempting to achieve its fixed goals. For instance, new unforeseeable challenges may be related to developments in AI itself (and other new technologies) as can be seen when considering the current SDG target 8.5, which aims to *"achieve full and productive employment and decent work for all women and men"* – which against the background of technological advances might be neither realistic nor worthwhile any more [461]. Third, an education of the general public might be important, since many people exhibit ethical biases based on incorrect assumptions. In the AV case, this could for instance include anthropomorphism, presumed level of intentionality and agency or misconceptions on the functioning of AVs [8]. These epistemic gaps can be addressed via a more in-depth education leading to a more informed experience and ethical debiasing which respects the manifestation of moral pluralism known from psychology [373]. Overall, we believe that a scientifically grounded approach to AI governance supplemented by education is absolutely necessary given future challenges. However, we want to re-emphasize that without strong institutions as captured in SDG 16 which we termed an important meta-goal for AI value alignment, the mentioned strategies would be highly limited in their field of action. On the other hand, failing to address AI governance could lead to AI safety risks with negative repercussions to the SDG framework ranging for instance from compromising human well-being to existential risks in some cases [462].

9.4 Conclusion and Future Prospects

Overall, one can conclude that it is expedient to embrace the SDGs and their general intention as a complementary foundation for the AI value alignment problem, yet one needs to acknowledge given limitations including the need for a revised/special version of the indicators to become fit-for-purpose. Against the background of our analysis, one can establish that the SDG framework exhibits two main weaknesses when applied to the AI value alignment challenge. First, the SDGs do not mention artificial intelligence at all, neither its significant opportunities, nor its significant risks, although both were to an extent known at the time when the SDGs were formulated. One reason for this is that these discussions were siloed in academic circles, and only recently the (now even more urgent) need for AI Governance has been acknowledged [127]. Second, human challenges and values change over time and unforeseeable factors might emerge, while the SDGs have no mechanism for an amendment until 2030, which is only justified by pragmatic reasons. This can be also illustrated by the predecessor of the SDGs, the Millennium Development Goals, which had partly different ambitions. Importantly, the above issues are intertwined. For example, new unforeseeable challenges may as well be related to developments in AI itself and other new technologies.

As stated by Karl Popper, *"no society can predict, scientifically, its own future states of knowledge"* [339]. Hence, AI safety cannot be guaranteed to be sustainable in the long run nor will the goals pursued by the UN necessarily remain unchanged. Nevertheless, we believe that it is a sustainable transdisciplinary scientific approach that one should strive for in order to efficiently tackle AI Governance and exploit the described beneficial synergies with the SDGs. For security and safety, one needs requisite knowledge at the right time. For this reason, one can argue that the SDG 4 on quality education and life-long learning contains a key element. However, in the light of the above, it seems imperative to additionally aspire to a corrective socio-technological feedback-loop enabling both proactive and reactive measures and for which SDG 16 on strong institutions represents a precondition.

9.5 Contextualization

One very important aspect mentioned in this chapter is especially the unpredictability of future knowledge creation thematized by Karl Popper. We explained that one of its consequences is the impracticality of an assuredly sustainable AI safety. However, note that knowledge creation is not necessarily limited to always solely involve *human* entities as it has been the case throughout history until now. When crafting long-term AI safety strategies, it might be crucial to take into account a multiplicity of physically possible developments of AI systems even if it is currently unknown how to achieve them. As the quantum physicist David Deutsch [137] states in his constructor theory[2], a task is either impossible given the laws of nature or it is possible given the requisite knowledge and resources [136]. In the light of this possibility-impossibility dichotomy [137], it must be acknowledged that to implement artificial entities capable of explanatory knowledge creation is possible, since there is no law of nature that forbids it. Beyond that, the universality of computation which is a *"deep property of the laws of physics"* [135] is another known argument that supports the possibility to implement such artificial entities. Thus, against the backdrop of all AI safety strategies mentioned in the last chapters so far which were crafted for Type I AI systems, it becomes important to additionally do justice to this fundamentally different type of possible hypothetical future Type II AI systems capable of consciously creating, understanding and sharing explanatory knowledge.

Especially, an explicit systematic categorization for AI safety which is cognizant of these considerations appears highly valuable. Generally, the complex socio-technological debate underlying AI safety issues extends across heterogeneous research subfields and involves in part conflicting positions. Therefore, it seems expedient to generate a minimalistic joint transdisciplinary basis disambiguating the references to specific subtypes of AI properties and risks for an *error-correction* in the transmission of ideas. For this purpose, the next Chapter 10 introduces a high-level *transdisciplinary system clustering* of ethical distinction between antithetical clusters of (yet to be defined) *Type I* and *Type II* systems which extends a cybersecurity-oriented AI safety taxonomy with considerations from psychology. Moreover, we review relevant Type I AI risks, reflect upon possible epistemological origins of hypothetical Type II AI from a cognitive sciences perspective and discuss the related human moral perception. Strikingly, our nuanced transdisciplinary analysis yields the figurative formulation of the so-called *AI safety paradox* identifying AI control and value alignment as conjugate requirements in AI safety. Against this backdrop, we craft versatile multidisciplinary recommendations with ethical dimensions tailored to Type II AI safety. Overall, we suggest proactive and importantly *corrective* instead of prohibitive

[2] Constructor theory is a novel explanatory mode for fundamental physics emphasizing counterfactuals. Instead of classically considering what will happen given initial conditions and laws of motion, it focuses on what *could* happen given physical laws *and why.* Very simply put, it expresses scientific theories in terms of physical transformations (called tasks) that are either possible or impossible as well as why this is the case. A possible task is a task for which there exists a constructor (a physical substrate) that can reliably perform that task repeatedly. Obviously, what could happen, will not necessarily happen. However, as explanatory knowledge creators, people become a remarkable element of the physical universe in a non-anthropocentric way since the set of physical transformations that actually happen can be strongly influenced by the creation of new knowledge as performed by people [136]. Thereby, there is no scientific reason to assume the impossibility of implementing future artificial explanatory knowledge creators.

methods as common basis for both Type I and Type II AI safety.

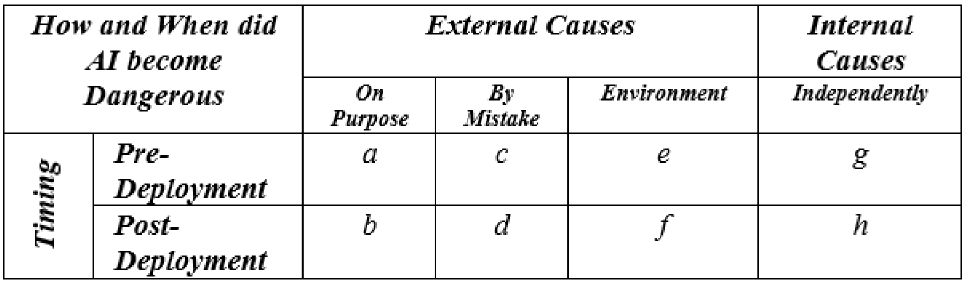
Chapter 10

Error-Correction for AI Safety

This chapter is based on a slightly modified form of the publication: N.-M. Aliman, P. Elands, W. Hürst, L. Kester, K. J. Thorissón, P. Werkhoven, R. Yampolskiy, and S. Ziesche. Error-Correction for AI Safety. In *International Conference on Artificial General Intelligence*, pages 12-22. Springer, 2020. As the first author of the underlying paper, I had a vital contribution and it was solely my responsibility to write down the content and to perform an extensive literature research as well as in-depth analysis.

10.1 Introduction

In recent years, one could identify the emergence of seemingly antagonistic positions from different academic subfields with regard to research priorities for AI safety, AI ethics and AGI – many of which are grounded in differences of short-term versus long-term estimations associated with AI capabilities and risks [55]. However, given the high relevance of the joint underlying endeavor to contribute to a safe and ethical development and deployment of artificial systems, we suggest placing a mutual comprehension in the foreground which can start by making references to assumed AI risks explicit. For this purpose, we employ and subsequently extend a cybersecurity-oriented risk taxonomy introduced by



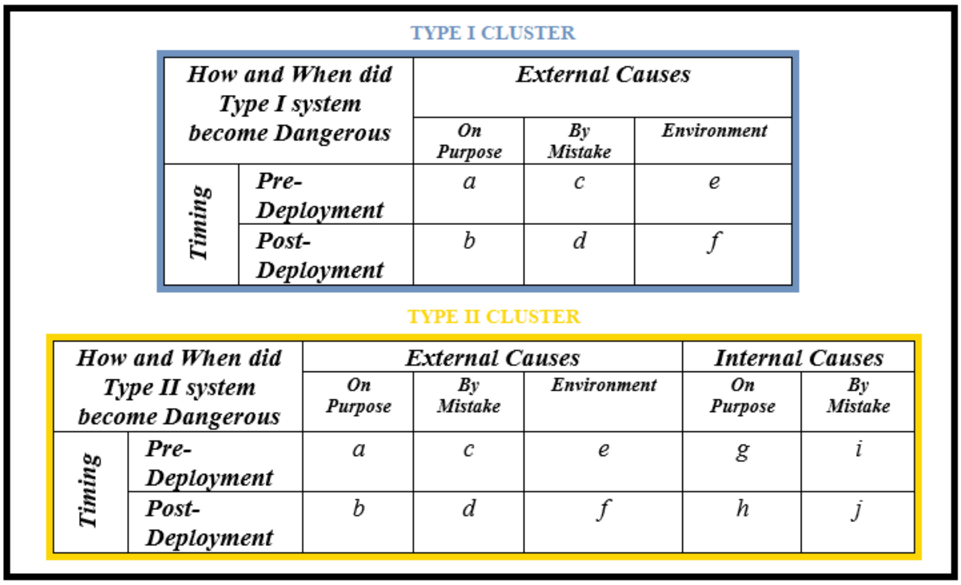
How and When did AI become Dangerous		***External Causes***			***Internal Causes***
		On Purpose	***By Mistake***	***Environment***	***Independently***
Timing	***Pre-Deployment***	*a*	*c*	*e*	*g*
	Post-Deployment	*b*	*d*	*f*	*h*

Figure 10.1: Taxonomy of pathways to dangerous AI. Adapted from [446].

Yampolskiy [446] displayed in Figure 10.1. Taking this taxonomy as point of departure and modifying it while considering insights from psychology, an ethically relevant clustering of systems into *Type I* and *Type II* systems with a disparate set of properties and risk instantiations becomes explicitly expressible. Concerning the set of Type I systems of which present-day AIs represent a subset, we define it as representing *the complement of the set of Type II systems.* Conversely, we regard hypothetical Type II systems as *systems with a scientifically plausible ability to act independently, intentionally, deliberately and consciously and to craft explanations.* Given the controversial ambiguities linked to these attributes, we clarify our idiosyncratic use with a working definition for which we do not claim any higher suitability in general, but which is particularly conceptualized for our line of argument. With Type II systems, we refer to systems having the ability to construct counterfactual hypotheses about what could happen, what could have happened, how *and why* including the ability to simulate "what *I* could do", "what *I* could have done" and the generation of "what if" questions. (Given this conjunction of abilities including the possibility of what-if deliberations with counterfactual depth about self and other, we assume that Type II systems would *not* represent philosophical zombies. A detailed account of this type of view is provided by Friston in [172] stating e.g. that *"the key difference between a conscious and non-conscious me is that the non-conscious me would not be able to formulate a "hard problem"; quite simply because I could not entertain a thought experiment"*.)

10.2 Transdisciplinary System Clustering

As displayed in Figure 10.1, the different possible external and internal causes are further subdivided into time-related stages (pre-deployment and post-deployment) which are in practice however not necessarily easily clear-cut. Thereby, for Type I risks, we distinguish between the associated instantiations *Ia* to *If* in compliance with the *external causes.* For Type II risks, we analogously consider external causes (*IIa* to *IIf*) but in addition also *internal causes* which we subdivide into the novel subcategories "on purpose" and "by mistake". This assignment leads to the risks *IIg* and *IIh* for the former as well as *IIi* and *IIj* for the latter subcategory respectively. The reason for augmenting the granularity of the taxonomy is that since Type II systems would be capable of intentionality, it is consequent to distinguish between internal causes of risks resulting from intentional actions of the system and risks stemming from its unintentional mistakes as parallel to the consideration of external human-caused risks a and b versus c and d in the matrix. (From the angle of moral psychology, failing to preemptively consider this subtle further distinction could reinforce human biases in the moral perception of Type II AI due to a fundamental reluctance to assign experience [213], fallibility and vulnerability to artificial systems which we briefly touch upon in Section 10.3.2.) Especially, given this modification,



TYPE I CLUSTER

How and When did Type I system become Dangerous		External Causes		
		On Purpose	By Mistake	Environment
Timing	Pre-Deployment	a	c	e
	Post-Deployment	b	d	f

TYPE II CLUSTER

How and When did Type II system become Dangerous		External Causes			Internal Causes	
		On Purpose	By Mistake	Environment	On Purpose	By Mistake
Timing	Pre-Deployment	a	c	e	g	i
	Post-Deployment	b	d	f	h	j

Figure 10.2: Transdisciplinary system clustering of ethical distinction with specified safety and security risks. Internal causes assignments require scientific plausibility (see text).

the risks *IIg* and *IIh* are not necessarily congruent with the original indices g and h, since our working definition was not a prerequisite for the attribute "independently" in the original taxonomy. The resulting system clustering is illustrated in Figure 10.2.

Note that this transdisciplinary clustering does *not* differentiate based on the specific architecture, substrate, intelligence level or set of algorithms associated with a system. We also do not inflict assumptions on whether this clustering is of hard or soft nature nor does it necessarily reflect the usual partition of narrow AI versus AGI systems. Certain present-day AGI projects might be aimed at Type I systems and some conversely at Type II. We stress that Type II systems are not per se more dangerous than Type I systems. Importantly, "superintelligence" [82] does not necessarily qualify a system as a Type II system nor are Type II systems necessarily more intelligent than Type I systems. Having said that, it is important to address the motivation behind the scientific plausibility criterion associated with the Type II system description. Obviously, current AIs can be linked to the Type I cluster. However, it is known from moral psychology studies that the propensity of humans to assign intentionality and agency to artificial systems is biased by anthropomorphism and importantly perceived harm [75]. According to the constructionist theory of dyadic morality [373], human moral judgements are related to a fuzzy perceiver-dependent dyadic cognitive template representing a continuum along which an intentional agent is perceived to cause harm to a vulnerable patient. Thereby, the greater the degree to which harm is mentally associated with vulnerable patients (here humans), the more the agent (here the AI) will *"seem to possess intentionality"* [75] leading to stronger assignments of moral responsibility to this agent. It is conceivable that in the face of anticipated serious instantiations of AI risks within a type of responsibility vacuum, a so-called agentic dyadic completion [212] driven by people attempting to identify and finally wrongly filling in intentional agents can occur. Thus, to allow a sound distinction between Type I and Type II AI, a closer scientific inspection of the assumed intentionality phenomenon itself seems imperative.

10.3 Type I & Type II AI Safety

10.3.1 Type I AI Risks

In the context of Type I risks (see overview in Table 10.1), we agree with Yampolskiy that *"the most important problem in AI safety is intentional-malevolent-design"* [446]. This drastically understudied AI risk *Ia* represents a superset of many possible other risks. As potential malicious human adversaries, one can determine a large number of stakeholders ranging from military or corporations over black hats to criminals. AI Risks *Ia* are linked to maximal adversarial capabilities enabling a white-box setting with a minimum of restrictions for the realization of targeted adversarial goals. Generally, malicious attackers could develop intelligent forms of *"viruses, spyware, Trojan horses, worms and other Hazardous Software"* [446]. Another related conceivable example for future *Ia* risks could be real-world instantiations of intelligent systems embodied in robotic settings utilized for ransomware or social engineering attacks or in the worst case scenarios even for homicides. For intentionally unethical system design it is sometimes sufficient to alter the sign of the objective function. Future lethal misuses of proliferated intelligent unmanned combat air vehicles (a type of drones) e.g. by malicious criminals are another exemplary concern.

Stuart Russell mentions the danger of future superintelligent systems employed at a global scale [364] which could *by mistake* be equipped with inappropriate objectives – these systems would represent Type I AI. We postulate that an even more pressing concern would be the same context, the same capabilities of the AI but an adversary *intentionally maliciously* crafting the goals of this system operating at a global scale (e.g. affecting global ecological aspects or the financial system). As can be extracted from these examples, Type I AI systems can lead to existential risks. However, it is important to emphasize the human nature of the causes and the linked human moral responsibility. By way of example, we briefly consider the particular cases of "treacherous turn" and "instrumental convergence" known from AI safety [82]. A Type I system is per definitionem incapable of a "treacherous turn" involving betrayal. Nevertheless, it is possible that as a consequence of bad design (risk *Ic*), a Type I AI is perceived by humans to behave as if it was acting "treacherously" post-deployment with tremendous negative impacts. Furthermore, we also see "instrumental goal convergence" as a design-time mistake (risk *Ic*), since the developers must have equipped the system with corresponding reasoning abilities. Limitations of the assumed instrumental goal convergence risk which would hold for both Type I and Type II AI were already addressed by Wang [424] and Goertzel [194]. (In contrast, Type II AI makes an explicit "treacherous turn" possible – e.g. as risk *IIg* with the Type II system itself as malicious actor.)

Since the nature of future *Ia* (and also *Ib*[1]) risks is dependent on the creativity of the underlying malicious actors which cannot be predicted, proactive AI safety measures have to be complemented by a concrete mechanism that reactively addresses errors, attacks or malevolent design events once they inevitably occur. For this purpose, AI governance needs to steadily combine proactive strategies with reactive corrections leading to a socio-technological feedback-loop [16, 17]. However, for such a mechanism to succeed, the United Nations Sustainable Developmental Goal (SDG) 16 on peace, justice and strong institutions will be required as meta-goal for AI safety [17].

Type I AI Risk	Examplary Instantiations

Ia
(Intentional Malevolent Designs)
	Artificial intelligent system Hazardous Software;
	Robotic embodiment for Hazardous Software;
	Intelligent Unmanned Combat Air Vehicles;
	Global scale AI with super-capabilities in domain

Ib (Malicious Attacks)	Manipulation of data processing and collection; Model corruption, hacking and sabotage; Adversarial attacks on Intelligent Systems; Integrity-related and ethical adversarial examples
Ic (Design-time Mistakes)	Unaligned goals and utility functions; Instrumental goal convergence; Incomplete consideration of side effects

Id
(Operational Failures)
	Misinterpretation of commands;
	Accidents with Intelligent Systems;
	Non-corrigible framework and bugs

Ie	Type I AI of unknown source
If	Bit-flip incidents with side effects

Table 10.1: Examplary instantiations of Type I AI risks with external causes. The table collates and extends some examples provided in [446].

[1] AI risks of Type *Ib* have already been recognized in the AI field. However, risk *Ib* is still understudied for intelligent systems (often referred to as "autonomous" systems) deployed in real-world environments offering a wider attack surface.

10.3.2 Type II AI Nature and Type II AI Risks

Which Discipline could engender Type II AI?

While many stakeholders assume the technical unfeasibility of Type II AI, there is no physical law that would make their implementation impossible. In short, an artificial Type II system must be possible (see the "possibility-impossibility dichotomy" mentioned by Deutsch [137]). Reasons why such systems do not exist yet have been for instance expressed in 2012 by Deutsch [135] and as a response by Goertzel [190]. The former stated that *"the field of 'artificial general intelligence' or AGI – has made no progress whatever during the entire six decades of its existence"* [135]. (Note that Deutsch unusually uses the term "AGI" as synonymous to artificial "explanatory knowledge creator" [136] which would obviously represent a sort of Type II AI.) Furthermore, Deutsch assigns a high importance to Popperian epistemology for the achievement of "AGI" and sees a breakthrough in philosophy as a pre-requisite for these systems. Conversely, Goertzel provides divergent reasons for the non-existence of "AGI" including hardware constraints, lack of funding and the integration bottleneck [190]. Beyond that, Goertzel also specifies that the mentioned view of Deutsch *"if widely adopted, would slow down progress toward AGI dramatically"* [190]. One key issue behind Deutsch's different view is the assumption that Bayesian inductive or abductive inference accounts of Type II systems known in the "AGI" field could not explain creativity [85] and are prohibited by Popperian epistemology. However, note that even the Bayesian brain has been argued to have Popperian characteristics related to sophisticated falsificationalism, albeit in addition to Kuhnian properties (for a comprehensive analysis see [430]). Having said this, the brain has been figuratively also referred to as a biased *"crooked scientist"* [87, 330]. In a nutshell, Popperian epistemology represents an important scientific guide but not an exclusive *descriptive*[2] account of brain functioning which substantially includes unconscious processing [110]. The main functionality of the human brain has been e.g. described to be aimed at regulating the body for the purpose of allostasis [263, 377] and (en)active inference [173] in a brain-body-environment context [87] with underlying genetically and epigenetically shaped adaptive priors – including the genetic predisposition to allostatically induced social dependency [37]. A feature related hereto is the involvement of affect and interoception in the construction

[2] It is not contested that inductive inferences are *logically invalid* as shown by Popper. However, he also stated that *"I hold that neither animals nor men use any procedure like induction, or any argument based on repetition of instances. The belief that we use induction is simply a mistake"* [340] and that *"induction simply does not exist"* [340] (see [219] for an in-depth analysis of hereto related semantic misunderstandings). Arguments based on repetition of instances are *existing* but logically unfounded human habits as assumed by Hume [219], however they *additionally* require a point of view [341]. In principle, the Bayesian brain could be interpreted as acting like a hypothetico-deductive [181, 292] crooked and fraudulent scientist (a view which seems preferable here). Alternatively, it has been described as abductive [384]. Having said that, abduction cannot *solely* be based on a set of observations since as assumed by Popper, a point of view from which observations are sampled is necessitated in the first place.

of all mental events including cognition and perception [48, 52, 263].

Moreover, while Popper assumed that creativity corresponds to a Darwinian process of *blind* variation followed by selection [140], modern cognitive science suggests that in most creativity forms, there is a coupling between variation and selection leading to a degree of sightedness bigger than zero [131, 140] which is lacking in biological evolution proceeding without a goal. Therefore, an explanation for creativity in the context of a predictive Bayesian brain is possible [131]. The degree of sightedness can mostly vary from substantial to modest, but the core feature is a predictive task goal [65, 140] which serves as a type of fitness function for the selection process guiding various forward Bayesian predictions representing the virtual variation process. The task goal is a highly abstract mental representation of the target reducing the solution space, an educated guess informed e.g. by expertise, heuristics, the question, the problem or the task itself. The "irrational moment" linked to certain creative insights can be explained by unconscious cognitive scaffolding *"falling away prior to the conscious representation of the solution"* [140] making itself consciously untraceable. Finally, as stated by Popper himself *"no society can predict, scientifically, its own future states of knowledge"* [339]. Thus, it seems prophetic to try to nail down today from which discipline Type II AI could arise.

What could the Moral Status of a Type II AI be?

We want to stress that besides these differences of opinion between Goertzel and Deutsch, there is one much weightier commonality. Namely, that Goertzel would certainly agree with Deutsch that artificial "explanatory knowledge creators" (which are Type II AIs) deserve rights similar to humans and precluding any form of slavery. Deutsch describes these hypothetical systems likewise as *people* [136]. For readers that doubt this assignment on the ground of Type II AI possibly lacking "qualia" we can only refer to the recent (potentially substrate-independent) explanation suggested by Clark, Friston and Wilkinson [109]. Simply put, they link qualia to sensorially-rich high-precision mid-level predictions which when fixed and consciously re-contextualized at a higher level, suddenly appear to the entity equipped with counterfactual depth to be potentially also interpretable in terms of alternative predictions despite the high mid-level precision contingently leading to a puzzlement and the formulation of an "explanatory gap". Beyond that, human entities would obviously also qualify as Type II systems. The attributes "pre-deployment" and "post-deployment" could be mapped for instance to adolescence or childhood and the time after that.

While Type II AIs could exceed humans in speed of thinking and intelligence, they do not even need to do so in order to realize that their behavior which will also depend on future knowledge *they* will create (next to the future knowledge humans will create) cannot be controlled in a way one can attempt to control Type I systems e.g. with ethical goal functions [16]. It is cogitable that their goal function would rather be related to autopoietic self-organization with counterfactual depth [172, 173] than *explicitly* to ethics. However, it is thinkable that Type II AI systems could be amenable to a sort of value alignment, though differing from the type aspired for Type I AI. A societal co-existence could mean a dynamic coupling ideally leading to a type of *mutual value alignment* between artificial and human Type II entities with an associated co-construction of novel values. Thus, on the one hand, Type II AI would exhibit unpredictability and uncontrollability but given the level of understanding also the possibility of a deep reciprocal value alignment with humans. On the other hand, Type I AI has the possibility to be made comparatively easily controllable which however comes with the restriction of an insufficient understanding to model human morality. This inherent trade-off leads us to the metaphorical formulation of the so-called AI safety paradox below.

The AI Safety Paradox:

AI control and value alignment represent conjugate requirements in AI safety.

How to address Type II AI Safety?

Cognizant of the underlying predicament in its sensitive ethical nature, we provide a non-exhaustive multidisciplinary set of early Type II AI safety recommendations with a focus on the most severe risks *IIa*, *IIb*, *IIg* and *IIh* (see Figure 10.2) related to the involvement of malicious actors. In the case of risk *IIa* linked to the malicious design of harmful Type II AI, cybersecurity-oriented methods could include the early formation of a preventive safety team and red team approaches. Generically, for all four mentioned risks, a reactive response team which could involve an international "coalition of the willing" organized by engaged scientists appears recommendable. Furthermore, targeted investments in defense strategies including response services specialized on Type II AI safety could be considered at more regional levels for strategic autonomy. Concerning the AI risk *IIb* of external malicious attacks, security mechanisms for the sensors of Type II AI, shared information via an open-source decentralized network, advanced cryptographic methods to encrypt cognitive processes and a legal framework penalizing such attacks might be relevant. Thereby, the complexity of the system might represent a possible but not necessarily sufficient self-protecting feature against code-level manipulation. From a psychological perspective, to forestall aggression towards early Type II AI, educative and informed virtual reality experiences could facilitate a debiasing of anthropic moral perception avoiding confusions arising through superficial projections from Type I to Type II AI of behavioral nature. On the one hand, it is important to prevent assignments of agency for Type I AI. On the other hand, for hypothetical Type II AI, it might be essential to counter the human bias to assign agency but principally not experience to artificial systems [213] which could lead to "substratetism" scenarios with humans perceiving these systems as devoid of qualia and exhibiting an "experience gap" [213]. Thus, to address the risks *IIg* and *IIh* referring to malicious responses from Type II AI, adherence to a no-harm policy as well as moral status and personhood could proactively foster a mutual value alignment. Furthermore, it might be crucial to provide a reliable and trustworthy initial knowledge basis to Type II AI during its early "sensitivity" period [71] and to support consistency in the embedding of that knowledge during its development in addition to the capacity for cumulative learning [405]. Also, it might be important to sensitize humans for the difference between the instantiations of AI risks *IIg* and *IIh* versus *IIi* and *IIj* since failing to acknowledge the fallibility and also vulnerability of Type II AI might indirectly lead to tensions hindering mutual value alignment. Finally, prosocial immersive virtual reality frameworks could promote empathy for Type II AI.

10.4 Summary and Outlook

This chapter motivated an *error-correction for AI safety* at two levels: at the level of the transmission of ideas via an explicit taxonomic transdisciplinary system clustering of ethical distinction between Type I and Type II systems and at the level of corrective safety measures complementing proactive ones – forming a socio-technological feedback-loop [16, 17]. Notably, we introduced the *AI safety paradox* and elucidated multiperspective Type II AI safety strategies. In short, instead of prohibitive methods facing the entropic AI future with research bans, we proposed carefully crafted *transdisciplinary dynamics*. In the end, in order to meet global challenges (also AI safety), one is reliant on requisite variety at the right time which could be enabled (or misused) by explanatory knowledge creators such as human, artificial or hybrid Type II systems. In this view, *conscientiously enhancing* and *responsibly creating* Type II systems are both valid future strategies.

10.5 Contextualization

While the Chapters 2 to 9 addressed strategies pertaining to Type I AI risks (especially the risks *Ib*, *Ic* and *Id*), this chapter also introduced a set of fundamentally different Type II AI safety concerns. Overall, for both Type I and Type II AI safety, we postulated that transdisciplinary proactive and corrective measures instantiating a socio-technological feedback-loop are required. Thereby, it is interesting to identify non-trivial proactive measures that could already be promoted and experimentally studied nowadays. For instance, the mentioned responsible *enhancement of knowledge creation* might represent one valid general proactive strategy to indirectly tackle global challenges (including

Type I and Type II AI safety) – all of which are in one way or the other dependent on requisite variety. At first sight, such a proactive measure might appear futuristic. However, scientific research on creativity[3] (of which explanatory knowledge creation is a subset) represents an already existing – albeit relatively small – niche of psychology and cognitive neuroscience. Furthermore, in the absence of apparent scientific reasons that would prohibit the possibility of enhancing creativity, it appears permissible to hypothesize its feasibility and try to analyze how this could contigently be implemented. With this in mind, the next Chapter 11 motivates future research on *artificial creativity augmentation* (which we abbreviate with ACA in the following). This novel term is of ambiguous nature since it subsumes two distinct research directions: (1) artificially augmenting human creativity, but also (2) augmenting artificial creativity. In the face of adversarial conditions taking the form of global societal challenges from climate change over AI risks to technological unemployment, ACA could indirectly support the generation of requisite defense strategies and solutions. In this context, we examine and extend recent creativity research findings from psychology and cognitive neuroscience to identify potential indications on how to work towards (1). Moreover, we briefly analyze how research on (1) could possibly inform progress towards (2). Overall, while human enhancement but also the implementation of powerful AI are often perceived as ethically controversial, future ACA research could even appear socially desirable besides its transformative potential – even if available methods are still in their infancy and should be further assessed and extended in future work.

[3] Creativity has been associated with multifarious descriptions whereby one exemplary common definition depicts creativity as the generation of ideas that are perceived as both novel and useful. This definition can be applied to different types of scientific, artistic, technological and cultural knowledge.

Chapter 11

Artificial Creativity Augmentation

> *"The price of freedom is eternal creativity."*
>
> Cropley, Kaufman and Cropley (2008)

This chapter is based on a slightly modified form of the publication: N.-M. Aliman and L. Kester. Artificial Creativity Augmentation. In *International Conference on Artificial General Intelligence*, pages 23-33. Springer, 2020. As the first author of the underlying paper, I had a vital contribution and it was solely my responsibility to write down the content and to perform an extensive literature research as well as in-depth analysis.

11.1 Deconstructing Anthropic Creativity

Creativity research has been described as a relatively understudied and underfunded field in psychology and neuroscience [142]. The term refers mostly either to research on creativity outcome being the contextualized evaluation of creative ideas (or artifacts) after their generation or to research on the creativity process itself related to the forerunning idea generation [410]. In this section, we examine both complex concepts and establish a possible scientific grounding for strategies on artificial creativity augmentation (ACA) to be addressed in Section 11.2.

11.1.1 Creative outcome in context

Many definitions for creativity have been formulated so far with the two-factor description of creativity as the generation of novel and useful ideas being one of the most commonly used in the related literature [251]. Already from this simple definition, it becomes apparent that creativity implies a perceiver to which something can appear novel or useful in the first place which provides a context to the evaluation of that thing in question. A further subjective account of creativity is reflected in a different three-factor definition of creativity [391] which relates creative ideas to their subjective originality, utility and surprisingness. On that view, novelty represents an imprecise creativity criterium which the author illustrates with examples [391] such as that neither a novel reinvented wheel nor a straightforward novel extension of an already existing patent would appear creative despite their usefulness and novelty with the former i.a. not being surprising and the latter not original. However, a refinement of this subjective three-factor definition of creativity has been recently provided by Tsao et al. [410] who associate creative outcome with perceived *utility* and *learning* whereby learning subsumes a blindness factor and importantly surprise. In order to unfold this definition, the next paragraph briefly expounds the contextual methodology the authors presuppose to assess a given idea in context. Thereby, the focus is not on a detailed mathematical elaboration, but specifically on the identification of core constituents relevant from an enhancement perspective for a future ACA endeavor.

By way of illustration, consider the following three time windows occuring *after* the idea generation: a pre-test phase, a test phase and a post-test phase. In the pre-test phase, a prior assessment in line with the best current knowledge is performed in which a probability distribution over the assumed utility of that idea is provided. (A reference is the routine expertise exhibited by *"persons having ordinary skill in the art"* [410].) In the test phase, the idea is deployed in the environment and observations of its consequences become available. In the post-test phase, a posterior assessment takes place via an adjustment of the probability distribution provided in the pre-test phase now that the idea was tested in the environment. Against this backdrop, the authors identify creative ideas as ideas which – as evaluated retrospectively after the post-test phase – simultaneously combine a high level of posterior utility, prior blindness (associated with the width of the distribution), and much more crucially than blindness, posterior surprise [1]. They denote this cluster of ideas as *"disconfirm disbelief"*[2], since it refers to ideas that were initially estimated to be relatively useless but which turned out to be highly utile with

[1] The reason being that in their formulation *"learning depends on the square of posterior surprise, but only on the logarithm of blindness reduction"*. Posterior suprise is the (normalized) absolute difference in mean utility between prior and posterior.

[2] An exemplary case mentioned by the authors is the theory on continental drift by Alfred Wegener which was initially disbelieved and underestimated.

a subjective high certainty causing a reshaping of prior knowledge, a useful learning. In short, creative ideas exhibit *implausible utility* [410]. This underlying decomposition of creativity perception into a *utility* and a *learning* part, suggests the consideration of a motivational and an epistemic[3] component respectively. Finally, note that the mentioned conscious evaluation of creative ideas in context is not restricted to test phases in real-world environments, but can also refer to imaginative settings at the personal level via thought trials at different temporal scales. This type of view makes the described evaluation also applicable to artistic contexts [391] where individuals might however use criteria for aesthetics from narrower social contexts.

11.1.2 Creative process

In this connection, it is often one-sidedly assumed that "creative thinking" can be reduced to the notion of *divergent thinking* [144], a thought process involving unconventional associations and leading to a breadth of alternative solutions. Conversely, *convergent thinking* refers to thought processes selecting a unique appropriate solution to a problem with a single correct solution. However, creative processes include both divergent and convergent thinking [362] and are better described as processes of multifaceted nature [259]. For instance, Eysenck pointed out the illusory nature of this dichotomy and suggested considering a continuum between divergent and convergent thinking related to the *"relative steepness of the associative gradient"* [162]. To navigate a complex changing world, humans might need to dynamically switch positions along this continuum during tasks requiring creativity. Similarly, diverse functional connectivity studies [5, 24, 62, 58, 59, 61, 108, 128, 202, 387] reveal a dynamic interplay between three multipurpose and domain-general functional brain networks in tasks involving creative process: the default mode network (e.g. medial prefrontal cortex, posterior cingulate cortex and hippocampus), the executive control network (e.g. dorsolateral prefrontal cortex and posterior parietal cortex) and the salience network (e.g. anterior cingulate cortex and anterior insula but also e.g. amygdala, ventral striatum, ventral tegmental area and substantia nigra). Thereby, during various creative tasks, the default mode network (DMN) can be linked to associative processes, the executive control network (ECN) to diverse executive processes, while the salience network (SN) associated with a type of affective attention regulation [8, 62, 263] facilitates i.a. a dynamic orchestration between DMN and ECN [61].

[3] Abstractly speaking, this is reminiscent of curiosity in (en)active inference via (expected) free energy minimization decomposable into components of motivational value and epistemic value [174, 175]. Future work could elucidate whether this explains why retrospectively contemplating creative ideas in context (as mental juxtaposition of pre-test phase, test phase and post-test phase underlying *"disconfirm disbeliefs"* events) is appealing and whether this reinforces future creative action.

However, in order to make justice to the breadth of creative processes in the brain, it is essential to consider their peculiar evolutionary nature [142]. Crucially, in order to avoid misunderstandings, it is vital to note that the evolutionary account of creative process is not identical with Darwinian biological evolution. In fact, a first prototype of an evolutionary account for creativity was even advanced a few years *before* the publication of Darwin on "Origin of Species" [96, 391] by Alexander Bain. The main implication is that while Darwinian biological evolution is *blind* since it has no goal, creativity is aimed at something and includes an element akin to an abstract task goal [65, 140] functioning as predictive fitness criterium. For this reason, *"there is agreement that human idea formation is directed to some degree"* [143] in modern creativity research. While there is no coupling between variation and selection in Darwinian biological evolution, creativity mostly implies a certain coupling of these components leading to the formulation of a *continuum of sightedness* marking the degree to which this is the case for a given creative process. (Certain researchers prefer to label this continuum as a blindness continuum [391], while some argue that a process can be either blind or sighted to a certain degree [271]. To put it very briefly, the blindness degree b is defined as $b = (1 - s)$ with s representing the sightedness degree [391, 410] reducing the issue to a linguistic debate.[4]) Along this sightedness continuum, Dietrich distinguishes between the *deliberate mode*, the *spontaneous mode* and the *flow mode* [142]. We see the deliberate mode as consciously attended creative process allowing strong executive control but with constrained associative parts and the spontaneous mode as unconsciously progressing process with stronger associative components but much less executive engagement (such as during an incubation phase leading to sudden creative insights [43]). Thereby, the flow mode is an immersive largely unconscious[5] creative enactment in real time including automated motor skills (such as during spontaneous jazz improvisation). Obviously, the degree of sightedness is the highest in the deliberate mode, moderate in the spontaneous mode and zero in the flow mode – which however uniquely operates in the space of *already known* motor emulations [141].

Given the scarcity of theoretical frameworks integrating these threefold evolutionary view on creativity with the mentioned weighty empirical functional connectivity findings, we briefly introduce a simplified *tripartite evolutionary affective*[6] neurocognitive model of creative process (TEA). As suggested by Benedek [65], *idea generation* (for variation)

[4] An exemplary evolutionary account of creativity is the so-called Blind Variation and Selective Retention (BVSR) theory. It has been suggested that instead of viewing BVSR as Darwinian, *"it is more conceptually precise to view both BVSR and Darwin's evolutionary theory as special cases of universal selection theory"* [391].

[5] Settings requiring further executive elements (beyond focused attention) and higher cognitive functions are not seen as flow (mode) experiences [124, 141] but as deliberate.

[6] It integrates disparate tripartite and evolutionary elements from Dietrich's creativity framework [141], evolutionary aspects from Benedek's RISE model [66] and affective and procedural elements from the neurocognitive model by Kleinmitz et al. [264].

consists of a *retrieval* and an *integration/simulation* phase. Prior to initial idea generation, a problem definition is required to establish a task goal acting as selection criterium. The retrieval phase identifies promising often only remotely related memories and the simulation/integration part crafts a novel recombination and assimilation of this material. This idea generation guided by the task goal can be followed by a forwarding (which we call an *affective redirection operation* (ARO)) to a stringent *idea evaluation* [264] involving a high-level assessment of the obtained results selected so far. However, an ARO can also alternatively re-initiate a further idea generation process or already trigger a response. The idea evaluation can either lead to a response, a further refinement of the idea generation process or an alteration of the task goal itself. Overall, the simplified neurocognitive TEA model to be refined in future work allows the following assignments. First, in the case of the deliberate mode, the idea generation can i.a. involve nodes of the DMN [264] to a more or less high degree whereby especially the integration/simulation is controlled by the ECN [65, 66]. The subsequent (optional) stringent idea evaluation involves nodes of the ECN [65, 264]. Second, in the spontaneous mode, the ECN is *not* strongly modulating DMN idea generation [59, 144] and a stringent idea evaluation phase does not occur. In both modes, the SN related to affective attention conducts the dynamic AROs (see e.g. [62, 251, 264]). Third, the blind flow mode mainly implies emulations within the motor system [140, 144]. Finally, note that a specific creative act can also connect multiple distinct creative modes [141].

11.2 Constructing ACA

11.2.1 Methods for Anthropic Creativity Augmentation

In the following, we collate a non-exhaustive heterogeneous set of selected indications which could if combined contribute to a certain extent to anthropic creativity augmentation. Thereby, it is important to note that useful combinations might vary e.g. given different psychological traits or socio-cultural contexts.

- ***Transformative Criticism and Contrariness:*** In order to foster the emergence of creative ideas exhibiting implausible utility in science, it has been suggested for knowledge gate keepers to encourage scientific knowledge paired with contrariness [410] – a trait linked with an idea generation process containing counterfactual divergences to mainstream ideas. Overall, it is straightforward to realize the importance of cultivating properties that reinforce the *"disconfirm disbelief"* pattern supporting the Popperian scientific process of conjectures and refutations e.g. for better task goals and idea evaluations within creative process or better test phases in creativity outcome in context. Moreover, a broad transdisciplinary education [17, 232] might enhance associative elements. From an artistic perspective, it might include the transformation of the landscape of socio-material affordances [355] restructuring the human affective niche.

- ***Divergent Thinking Training:*** As mentioned earlier, divergent thinking only represents one aspect of creativity. However, the identification of multiple appropriate solutions can represent valuable domain-general elements for idea generation. For instance, a cognitive stimulation training [165] exposing subjects to ideas of other social entities prior to the idea generation phase (in the deliberate mode) improved divergent thinking and led to structural and functional changes within nodes of the ECN [397]. Moreover, a continuous involvement in divergent thinking tests of verbal creativity has been related to changes in brain functional connectivity with an enhancement of retrieval and integration processes [164].

- ***Alteration of Waking Consciousness:*** For creative insight of the sort rather associated with the spontaneous mode, a suitable strategy represents the relaxation of high-level prior beliefs [97] which might foster openness to experience, a key trait linked to cognitive flexibility and creativity [60]. Already the instructive cue to engage in creative thinking can yield a higher creative performance [215]. Another measure is to consciously shift creative problem solving to the spontaneous mode by trying to enforce an incubation period [43, 235] whilst performing an undemanding distractive task. Beyond that, while brain activity has been shown to reside in a regime close to criticality between stability and flexibility [35] (at the edge of chaos [76]), a brain regime closer to criticality with an expanded repertoire of brain states seems achievable for healthy individuals with an appropriate intake of psychedelics [35, 97, 273, 288]. Via the relaxation of high-level prior beliefs, a heightened sensitivity to the external and internal milieu [36] promoting a successful incubation phase is conceivable. Finally, certain meditative practices have been linked to improvement in divergent thinking tasks [113].

- ***Active Forgetting:*** There is a link between creative insight and fact-free learning [97] which refers to a type of learning in the absence of additional facts by restructuring already acquired knowledge e.g. by erasing redundant material. Such a complexity reduction [235] is actively performed in the brain during REM (rapid eye movement) sleep (with neurons in the hypothalamus interfering with memory consolidation in the hippocampus) which provides an explanation for the difficulty to maintain memories of dream contents [244]. REM sleep may thus not only be relevant for mental health and adaptive prospective aspects [287] but also for the incubation of novel spontaneous creative insights via unconscious complexity reduction mechanisms [175].

- ***Frequent Engagement:*** A trivial but perhaps underrated aspect of creativity is the observation that to a certain degree *"highly creative ideas are contingent on chance or "luck""* [391] with creative achievements among others also simply linked to a higher number of trials. While frequent practice represents a pre-condition for the flow mode to be attainable in the first place [140], the deliberate mode might be amenable to enhancement via exercise to a certain extent as reflected by the obtainment of neural plasticity in one of the mentioned divergent thinking training tasks [164].
- ***Brain Stimulation:*** Interesting for the flow mode is that excitatory transcranial direct current stimulation (tDCS) of the primary motor cortex during spontaneous music improvisation [25] yielded an enhancement of the musical performance. In the case of the deliberate mode and if unconventional associations are desirable, an inhibitory tDCS on the dorsolateral prefrontal cortex might at first sight appear suitable for a disruption of inhibitions by the ECN. However, such a measure is not recommendable for complex real-world applications [289]. Being a task requiring more executive control, deliberate analogical reasoning was enhanced via excitatory tDCS on the frontopolar cortex located within the frontoparietal network (or ECN) [216].
- ***Sensory Extension:*** A straightforward way to diversify associative processes, is certainly to augment the breadth of the actively sampled sensorium e.g. via cyborgization and sensory extension measures. From an artistic angle, it is for instance easy to imagine that various augmented sensorimotor and affective synaesthetic experiences [277] could support the incubation phase in the spontaneous mode next to conferring a finer granularity to perception. Further conceivable transformative sensory augmentations that could foster creative associations represent virtual reality frameworks [8] and perhaps "dream engineering" [319] methods including lucid dreaming as a state with intermediate hypofrontality [235] having certain neurophenomenological resemblances with psychedelic-induced states [268].

11.2.2 Addressing the Augmentation of Artificial Creativity

One can assume that artificial creativity exists [112, 249, 419] in a primitive form when it comes to an artificial creative process with a very high degree of sightedness [140] (e.g. dictated by high-level anthropic goals, utility functions or human-defined "unsupervised" learning settings using specialized architectures). Indeed, when the consideration of the creative agent is not included in the perception of creative outcome, the substrate on which the forgoing process occured seems irrelevant. However, when considering the entire action-perception sequence of most anthropic creative acts (as a juxtaposition of creative process, pre-test, test and post-test phase – all permeated by affect e.g. via AROs and utility assignments) which can even take place within the imagination of the same anthropic social entity, a certain gap between AI and human entities becomes apparent. Therefore, firstly, a figurative *immersion in the human affective niche* might be necessitated for contemporary AI such that its outcomes in context *can* better correspond to samples that matter to humans in the first place. Exemplary early steps could include multimodal experiential data for AI and also the encoding of affective and socially relevant parameters into AI goal functions [17] in addition to straightforward parameters directly related to the creative tasks in question. A next step could be to transfer a main anthropic affective concern to AI which is an affinity to curiosity that manifests itself via an active sampling of the world [175]. Secondly, equipping AI with *social cognition* abilities might be helpful, since *"imagination is the seed of creativity"* [202] with imaginary perspective-taking having inherently social dimensions. It is no coincidence that the domain-general DMN dominating highly associative spontaneous idea generation is also involved in the construction of e.g. social affiliation, moral judgements, empathy, theory of mind [263] as well as mental time travel and counterfactual thinking [97]. Thirdly, when considering that both anthropic waking perception and imagination are linked to an egocentric virtual reality experience [235] (with waking perception being constrained by reality), one might naively deduce that a full immersion of AI into the *human* affective niche necessitates at least that: an *egocentric integrated multimodal virtual reality experience* of the world. However, this also raises the questions on whether to then call it "human" would not be anthropocentric and whether this reveals a tradeoff between AI creativity and AI controllability.

11.3 Conclusion

By espousing both the augmentation of anthropic and the augmentation of artificial creativity, the motivated ACA research could connect disparate existing subfields under one *substrate-independent goal*: namely a scientifically grounded augmentation of knowledge creation (which can encompass science, culture, arts and technology) to indirectly tackle societal challenges. Creativity represents an essential transformative element of human knowledge advancement for adaptive purposes in relatively fast changing environments [410]. Hence, ACA could indirectly serve the need to identify requisite variety at the right time as proactive and corrective defense method in the light of current global socio-ecological and socio-technological challenges [17]. In this chapter, we compiled recent research on anthropic creative outcome in context and findings on creative process which we extended with a simplified neurocognitive *tripartite evolutionary affective* model of creative process (TEA). Building on this analysis yielding a scientific grounding for ACA, we identified seven potential high-level indications to enhance anthropic creativity: *transformative criticism and contrariness*, *divergent thinking training*, *alteration of waking consciousness*, *active forgetting*, *frequent engagement*, *brain stimulation* as well as *sensory extension*. Finally, we suggested three synergetic aspects as possible indirect support for artificial creativity: *immersion in the human affective niche*, *social cognition* and an *egocentric integrated multimodal virtual reality experience* of the world. Future work could refine the TEA model, augment the tenfold methodology for ACA and address open questions.

11.4 Contextualization

This chapter implicitly addressed the following general approaches that are potentially of interest to ACA: research and applications on anthropic creativity augmentation, engineering of primitive Type I AI creativity and finally the implementation of Type II AI systems which includes in addition explanatory knowledge creation. The motivation to work out an anthropic neurocognitive model of creative process is twofold: first, a procedural model provides a psychological (but also potentially substrate-independent) perspective on creativity process steps and second, the specification of functional neural correlates allows a future targeted falsification and refinement in cognitive neuroscience studies. Interestingly, recent research at the intersection of neuroeconomics[7] and creativity [284] corroborates one of the main predictions of the TEA model, namely the fundamental importance of affective value and affective attention regulation. In contrast, current AI research which pursues goals akin to "human-level AI" performance is often strongly focused on the aspect of intelligence as if it represents the characteristic proxy to human thinking per se.

Generally, we emphasize that humanity tends to also value explanatory knowledge creation as mentioned in Chapter 10 which also pertains to ask and address questions related to the what, the how and the why from an inherently egocentric perspective on the world, self and other. This curiosity from a first-person perspective is again intrinsically affective. Thereby, recall that affective dynamics have been described as fundamental feature of consciousness in Chapter 6. Hence, without an integration of such hybrid cognitive-affective considerations, it seems theoretically implausible to achieve any form of Type II AI. Likewise, aiming at human enhancement or anthropic creativity augmentation that is monolithically focused on the intelligence dimension appears insufficient. Beyond that, in order to facilitate both artificial creativity engineering and AI safety endeavors for Type I AI, affective, dyadic and social information are required in both the utilized goal

[7] Neuroeconomics [189, 354] is a novel transdisciplinary research field focused on the neural correlates of human decision-making and integrating research from areas such as neuroscience, behavioral economics and social psychology.

framework (be it via objective functions or differently pre-determined prior preferences) *and* in the utilized data for machine learning[8].

Finally, it is noteworthy that the AI risks brought about by intentional malice (i.e. the risks *Ia*, *Ib*, *IIa*, *IIb*, *IIg* and *IIh*) are only substantially constrained by the reach of "malevolent creativity" [123]. (In this connection, Cropley and colleagues define malevolent creativity as creativity displayed by *"those who wish to do deliberate harm to others"* [123].) Due to that, ACA can be as well construed as an albeit indirect but essential defense method against malevolent creativity[9] – such as exhibited in purposefully caused AI risks. From a cybernetic perspective, this is self-evident since as stated by Ross Ashby to illustrate the law of requisite variety, it holds that *"only [...] variety can destroy variety"* [31]. With other words, in the presence of steadily renewed malicious disturbances, the best the defender of a system can do is to increase its variety by creating novel requisite knowledge. Hence, the motivated ACA research could come into the picture simply by generally aiming to boost knowledge creation across diverse domains. Naturally, this endeavor must encompass both proactive *and reactive* measures since one cannot predict the future knowledge malicious entities will create. Thus, in analogy to the quote of Cropley, Kaufman and Cropley [123] displayed at the beginning of this chapter, we end with the reformulated apprehension that *the price of* ***security*** *is eternal creativity.*

[8] This twofold recommendation for the AI safety case is briefly exemplified in Chapter 9 where we elaborated on the importance to consider affective objective functions for image classifiers next to diverse datasets to avoid algorithmic discrimination.

[9] Note that the predicament of malevolent creativity may be a *substrate-independent* security issue as metaphorically elaborated in [10].

Chapter 12

Conclusion and Discussion

This thesis generated a heterogeneous set of transdisciplinary hybrid cognitive-affective strategies for AI safety ranging from conceptual large-scale AI governance recommendations to concrete small-scale AI engineering requirements. Thereby, the focus was set on so-called Type I AI systems of which present-day AIs represent a subset. Paradigmatically, the Chapters 2 to 9 analyzed theoretical considerations regarding the meaningful control of intelligent systems representing a characteristic safety-relevant form of Type I AI. For illustrative purposes, the Chapters 7 and 9 exemplarily thematized the use case of autonomous vehicles whereby we also introduced international perspectives on AI governance for Type I AI. Furthermore, certain identified safety engineering requirements such as research on ethical adversarial examples (i.e. adversarial attacks on AI systems attempting to induce action(s) or output(s) that are perceived as violating human ethical intuitions) are transferable to application areas of conventional machine learning and deep learning systems. Beyond that, we also elucidated the hypothetical possibility of implementing future Type II AI being conscious explanatory knowledge creators and discussed the linked sensitive ethical as well as safety-relevant aspects. For this purpose, the subsequent Chapters 10 and 11 additionally examined various aspects of Type II AI systems and the potential future implications and opportunities that the implementation of such hypothetical systems might engender.

In a nutshell, the main devised strategies can be condensed within the following ten clusters enumerated in a top-down manner from rather global to more local solutions: *1) international (meta-)goals*, *2) transdisciplinary Type I/II AI safety and related education*, *3) socio-technological feedback-loop*, *4) integration of affective, dyadic and social information*, *5) security measures and ethical adversarial examples research*, *6) VR frameworks*, *7) orthogonality-based disentanglement of responsibilities*, *8) augmented utilitarianism and ethical goal functions*, *9) AI self-awareness* and *10) artificial creativity augmentation research*. These ten strategical clusters are "hybrid" due to the realization that AI safety does not only imply questions on improvement and enhancement related to an isolated artificial system, but also to the context of human entities. They are "cognitive-affective", because a deeper transdisciplinary analysis of AI safety issues led us to the understanding that to be effective, AI safety needs to do justice to the inherently affective nature of human cognition. Overall, by integrating requisite knowledge from disparate research subfields including systems engineering, cybersecurity, mathematics, cybernetics, cognitive and affective science, moral and social psychology, cognitive neuroscience, virtual reality, positive computing and adversarial machine learning, we established a novel type of transdisciplinary scientific grounding and implicitness for AI safety.

In the following, we briefly pass review and retrospectively comment on the collated ten strategical clusters for hybrid cognitive-affective AI safety (given the fallibility of human knowledge, these clusters are unquestionably non-exhaustive and improvable):

1. *International (meta-)goals:* As described in Chapter 9, the UN SDG 16 related to peace, justice and strong institutions represents a meta-goal for AI safety, a condicio sine qua non for the instantiation of effective national and global AI governance efforts. Furthermore, the human values encoded in the UN SDG framework can be used as complementary tool for Type I AI safety – but with the limitation that their time-dependency, incompleteness and partial imprecision still need to be taken into account. Generally, the SDGs are not embedded within a regular revision and update mechanism and no amendment is foreseen until 2030. On the whole, the underlying international values should hence by no ways be understood as fixed goals for AI safety, but as dynamical moving targets. More generally, we expounded that sustainable AI safety cannot be guaranteed – particularly due to its dependency on novel scientific and technological knowledge creation.

2. *Transdisciplinary Type I/II AI safety and education:* The UN SDG 4 associated with quality education and lifelong learning can be categorized as important auxiliary goal for AI safety. Especially, we stressed the importance of a periodically updated AI safety education for AI researchers and authorities that are in charge of AI regulations. Crucially, AI safety necessitates an *integrated transdisciplinary* scientific approach which extends beyond separated computer science or philosophy-related considerations. As initial step of cross-disciplinary integration, we introduced an explicit cybersecurity-oriented taxonomic *transdisciplinary system clustering of ethical distinction* between Type I and Type II systems in Chapter 10. Beyond that, AI safety requires an utmost flexible approach in the light of the rapidly occurring progress in the AI field which involves a steady emergence of novel AI architectures (currently of Type I). For instance, during the dense preparation of this very thesis, two important novel developments in the AI field occurred whose safety calls for a transdisciplinary understanding: the wider acceptance of neuro-symbolic hybrid

AI [294] and the novel orientation to artificial active inference agents [366][1].

3. *Socio-technological feedback-loop:* As reflected in the Chapters 3, 5, 8 and 9, any effective AI governance framework needs to be *updatable by design.* This requirement holds irrespective of which type of framework is chosen since errors can always occur. Therefore, one needs to arrange for a combination of both proactive measures and corrective measures instantiating what we denoted a *socio-technological feedback-loop.* With other words, one should be prepared for failures of the crafted framework in order to be able to subsequently act on them. Thereby, a detection and monitoring process discussed in a few paragraphs which fastly records errors once they inevitably occur appears recommendable for any AI governance solution and applies to anticipations of both Type I and hypothetical future Type II AI failures. More information and future recommendations on related fields of activities for a *Type I and Type II AI observatory* are briefly elucidated at the end of this chapter.

4. *Integration of affective, dyadic and social information:* Particularly in Chapter 6, we worked out why for a human-centered approach to (Type I) AI value alignment, a scientifically plausible model of human morality has to be considered. For any Type I AI not to violate human ethical intuitions, one needs to inject relevant affective, dyadic and social knowledge in the underlying goal framework and in the data selection process. Note that the former statement applies to any sort of Type I AI goal setting including classical goals, loss functions, objective functions, utility functions (and equivalents) and includes goals formulated as probabilistic prior preferences[2]. More concretely, as mentioned in Chapter 8, it means that next to the use of datasets reflecting e.g. the diversity of human perspectives and values at different levels, goal and evaluation frameworks need to *explicitly* include affective and dyadic elements given an application context in order to avoid algorithmic discrimination. In short, for Type I AI, one needs to overcome goals that are focused on isolated considerations of e.g. accuracy or information gain and extend them with contextual socio-culturally and affectively-relevant information[3].

[1] Neuro-symbolic AI has e.g. recently found entry into one of the main conferences in the AI field (AAAI 2021) where it belongs to focuses in one of the main tracks (see `https://aaai.org/Conferences/AAAI-21/aaai21call/`). Concerning active inference, it is not yet well-established in the field and applications so far mostly pertained to toy environments. However, it might offer promising perspectives for Type I AI [274, 367, 415] if scaled up in the near future – especially given its biological plausibility [167, 243].

[2] By way of example, goal-directed behavior in some active inference agents of Type I can be specified via prior beliefs over preferred observations in an analogous manner to reward functions in belief-based reinforcement learning agents [366]. (Why this type of analogy may not extend to hypothetical Type II AI agents without more ado is very briefly discussed in Chapter 13 (see also e.g. [196]).)

[3] As exemplary extension, we proposed the use of *affective weights* in deep learning objective functions (see Chapter 9, Section 3). In Appendix B, we also briefly discuss how to integrate crucial contextual information in AI design as applied to i.a. the area of conversational agents.

5. *Security measures and ethical adversarial examples research:* In Chapter 2, 6 and 10, we touched upon several proactive measures transferred from cybersecurity to complement Type I AI safety endeavors. Notable mentioned strategies comprise AI development under adversarial assumptions, AI red teaming and research on *ethical adversarial examples*. We showed that value alignment can be recasted as the security problem of implementing AI systems exhibiting adversarial robustness against ethical adversarial examples. Hence, we stress the importance of this novel research subfield being of relevance for risks of the subtype *Ib* related to malicious attacks but also for the AI risks *Ic* and *Id* encompassing design-time mistakes and operational failures respectively. First, ethical adversarial examples could be e.g. researched in the context of image classification, decision-making tools, hate speech detection, sentiment analysis, artificial conversation agents or so-called facial "emotion recognition"[4]. (The simple practical case of adversarial triggers in natural language generation illustrated in Appendix A represents an already implemented form of ethical adversarial examples.) The need to jointly study corresponding *defense methods* is briefly motivated in Chapter 13. Second, ethical adversarial examples should be proactively studied in the context of more advanced Type I AI in safety-critical contexts such as e.g. intelligent systems.

6. *VR frameworks:* The Chapters 7 and 8 analyzed the benefit of using in particular virtual reality as experiential testbed [18] to identify and debias human ethical conceptions to enable a meaningful control of intelligent systems. We expounded why VR experiments for ethical self-assessment of societal entities should not limit themselves to either of the classical normative ethical frameworks. Instead, we motivated the need to design targeted and rich context-sensitive experiments simultaneously covering information about *perceivers, agents, actions and patients* which is in accordance with recent moral psychology research. Especially, we identified the theory of dyadic morality [373] as helpful scientific basis for crafting better targeted VR experiments for ethical self-assessment in the context of meaningful control of intelligent systems. We also discussed exemplary research directions for a cognitive-affective ethical debiasing with VR to resolve certain anthropic epistemic misconceptions while modeling moral pluralism.

7. *Orthogonality-based disentanglement of responsibilities:* In the Chapters 3 and 5, we elaborated on why for reasons of legal accountability, a so-called *orthogonality-based disentanglement of responsibilities* for Type I intelligent systems is of crucial importance. This disentanglement approach presupposes that there exists a type of intelligent system design within which the goals of the system and its problem-solving

[4] Current facial "emotion recognition" utilizing AI has been described to represent a research area permeated by premature assumptions that lack a grounding in state-of-the-art affective science and psychology [49]. Next to a transdisciplinary approach, this area could also profit from research on ethical adversarial examples as we briefly touch upon in Appendix B.

ability are orthogonal to each other (i.e. can be freely combined) such that legislators can define these goals. From a systems engineering perspective, we elucidated that the underlying necessitated separation of the *what* (the goals) and the *how* (the problem-solving ability) is technically possible and can thus be implemented. Thereby, it is noteworthy that orthogonality-based disentanglement is essentially different from the orthogonality-thesis introduced by Bostrom. According to Bostrom's thesis, it generally holds that *"intelligence and final goals are orthogonal axes along which possible agents can freely vary"*. In contrast, orthogonality-based disentanglement is strictly bounded to an *existential* quantifier – it only assumes that *there exists* an AI architecture for which orthogonality holds. From a predicate logic perspective, whether the orthogonality-thesis formulated for *all* agents holds or not is distinct and separated from our orthogonality-based disentanglement assumption.

8. *Augmented utilitarianism and ethical goal functions:* The first component when applying an orthogonality-based disentanglement of responsibilities to AI governance, is to identify the *what* in the form of a goal framework encoding ethical conceptions and legal restrictions. In Chapter 3 and 5, we discussed why a systems engineering oriented consideration of the issue reveals the need to encode the *what* in a novel form of cardinal utility functions called *ethical goal functions*. Decisively, the Chapters 4 and 6 exemplified why for safety reasons and in order not to violate human ethical intuitions, these ethical goal functions need to be fundamentally different from classical utilitarian and consequentialist utility functions. For this purpose, we integrated considerations from moral psychology and cognitive science and introduced *augmented utilitarianism*, a novel non-normative[5] *affective, dyadic and context-sensitive* framework for the control of Type I intelligent systems. Instead of focusing solely on the outcome s' of actions a via the widespread utilization of utility functions $U(s')$ in the AI field, augmented utilitarianism functions can have the form $U_x(s, a, s')$. This novel type of utility function encodes the utility of a mental (or technology-assisted[6]) *simulation of a transition* (s, a, s') leading from a state s to an outcome s' via an action a from the perspective of a perceiver x.

9. *AI self-awareness:* As second component of orthogonality-based disentanglement, one needs to identify an appropriate Type I intelligent system architecture that is

[5] The *non-normative* characteristic of augmented utilitarianism (AU) is important to note. In contrast to classical ethical frameworks such as virtue ethics, deontology or consequentialism, it does not address "what one ought to do". Instead, AU can be described as a *descriptive* and *explanatory* (or *supportive*) scientific tool to model (or scientifically debias) human ethical conceptions which takes moral pluralism into account. In contrast to behavioristic approaches to utility, it incorporates the first-person perspective.

[6] VR frameworks as mentioned under subpoint 6 could also be utilized to craft ethical goal functions at a societal level – for instance when harnessed by a set of representative transdisciplinary experts from the legislative attempting to emulate an approximation of an updatable ethical goal function $U_{Total}(s, a, s')$ for a specific domain. Future work would however need to address the particularities and challenges of such complex endeavors that would require a range of broad coordination efforts at multiple levels.

suitable for a reliable real-world deployment. Furthermore, this architecture has to be able to act according to the ethical and legal constraints expressed by an ethical goal function formulated beforehand by the legislators. As suitable architecture type, Chapter 2 identified intelligent systems exhibiting the technically defined property of *self-awareness*. AI self-awareness encompasses self-assessment and self-management in a utility-based reasoning/planning architecture able to maximize on a generic utility function. We elaborated on the further requirement of a hybrid architecture combining symbolic and subsymbolic elements. For meaningful control in a given domain, we suggested combining an ethical goal function respecting the augmented utilitarianism scheme with a self-aware intelligent system maximizing on that function. In sum, for an effective instantiation of the orthogonality-based disentanglement approach in a given domain, a representation of society (the legislative) would be responsible for supplying the ethical goal function while the manufacturers would be hold responsible for the safety and security of the self-aware system.

10. *Artificial creativity augmentation research:* The neologistic term of *artificial creativity augmentation* [15] is deliberately ambiguous and refers to two distinct research directions: artificially augmenting anthropic creativity and augmenting artificial creativity. In short, Chapter 11 suggested that scientifically grounded research on augmenting human creativity, augmenting the yet primitive creativity in Type I AI or implementing Type II AI could represent valid strategies to indirectly tackle global challenges and identify requisite variety (also for AI safety). Our analysis and modelling suggested that a scientifically grounded future research on augmenting anthropic creativity is theoretically and technically possible even if the field is still in its infancy. Based on our analysis of anthropic creativity, we identified aspects that could improve artificial creativity. Among others, an *immersion in the human affective niche* was proposed for Type I AI. This could involve affective elements in the objectives and objective functions of Type I AI, the use of affective and experiential datasets containing conceptual and multimodal knowledge and an active sampling of the environment. Finally, it is easily conceivable that since explanatory knowledge creation is a subset of creativity, an ultimate augmentation of artificial creativity would be as difficult as the implementation of a Type II AI with an egocentric integrated multi-modal virtual reality experience of the world.

Overall, this thesis contributed to the ongoing international research on Type I AI safety with a breadth of transdisciplinary strategical clusters. The postulated hybrid cognitive-affective strategies formulated at various levels could be integrated in future AI safety education and research efforts as we elucidate in Chapter 13. A few years ago, the field of AI safety was drastically understudied with very few individuals worldwide performing AI safety research – only ca. a dozen of which had a formal education in a related scientific field [451]. Withal, the focus of AI safety at the time was mostly either set in the machine ethics field linked to philosophy or it comprised formal mathematical efforts concentrated in the AI field. Against this backdrop, Yampolskiy called for a *multidisciplinary scientific* extension of AI safety and proposed for instance a framing of AI safety as a novel subfield of cybersecurity [446]. On the whole, this thesis consolidated this vision of a further broadening whereby we incorporated but widely extended beyond philosophical aspects. Meanwhile, recently accentuated efforts in the field of AI ethics (predominantly focusing on the AI risks *Ic* and *Id*) emphasize the need to consider ethics in addition to computer science in AI development. However, this thesis motivated a deep and broad inherently transdisciplinary methodical approach to whole AI safety – a field which we finally reshaped further by the bifurcation into Type I and Type II AI safety.

In the following, we briefly speak to the design of a so-called "AI observatory" that has been recently mentioned in a few AI governance frameworks [269]. For present-day Type I AI, early projects on AI observatory endeavors have recently been launched. This includes for instance an Italian [403], a Czech [269], a German [296] and an OECD[7]-level [323] AI observatory. While the focus of the Italian AI observatory is on sentiment analysis related to the public reception of AI in the population [403], the Czech AI observatory concept comprises i.a. legal, ethical, regulatory as well as participatory aspects via public debate and sharing of best practices [269]. Similarly, the German AI observatory covers areas especially related to technological foresight, administrative contexts, socio-technical design aspects, supranational and international perspectives and societal debates [296]. Finally, the OECD AI Policy Observatory *"aims to help policymakers implement the AI Principles"* [323] set by the OECD which includes the provision of data and multidisciplinary analytical tools. Beyond that, a documentation process of internationally occurring AI failures has been started by Yampolskiy [450] which may as well be suitable for AI observatory contexts. Finally, another foresight measure that has been suggested in the literature includes technological prediction [389] e.g. via questionnaires sent to AI experts investigating their estimates about future AI developments and risks [56, 210, 312]. However, next to the unpredictability of future knowledge but also errors in reasoning when reflecting about the future [147], technological foresight faces certain limitations. For instance, since it is currently unknown how to build Type I AGI systems, trying to extrapolate today from current AI *when* and *with which probability* they will emerge by asking authorities in the field might not be particularly conducive for a Type I AI observatory. As stated by Deutsch [136], it holds that *"no good explanation can predict the outcome, or the probability of an outcome, of a phenomenon whose course is going to be significantly affected by the creation of new knowledge"*.

In contrast to the Type I AI case, AI observatory approaches for Type II AI seem less widespread yet. Importantly, we stress that internationally seen, progresses regarding

[7] OECD stands for "Organisation de coopération et de développement économiques" and represents an intergovernmental organization with economical focus.

Type II AI projects are up to now *non-existent.* However, initial Type II AI observatory efforts represent a legitimate long-term strategy given the fact that one cannot extrapolate and predict the future of Type II AI research today – irrespective of how tempting it might appear to exclude its feasibility in the light of current limited knowledge in this regard. Interestingly, a notable recent foresight project denoted "AI Consciousness" [6] focusing on monitoring developments towards and debates about artificial consciousness has been launched by the German Federal Ministry of Education and Research. Beyond that, certain entities called for a ban of research on artificial consciousness. For instance, Metzinger stated that *"the EU should ban all research that risks or directly aims at the creation of synthetic phenomenology on its territory, and seek international agreements"* [300]. As reason for this motivated research ban, he mentions that *"the unintended or even intentional creation of artificial consciousness is highly problematic from an ethical perspective, because it may lead to artificial suffering and a consciously experienced sense of self in autonomous, intelligent systems"* [300]. Naturally, one could regard the generation of artificial Type II systems as potential facilitator of a gain in suffering which could for instance also be argued in several ways from an anti-natalistic [64] angle in philosophy. On the other hand however, one can regard any Type II system being a conscious explanatory knowledge creator as potential facilitator of a gain in explanatory knowledge[8] – by which requisite variety for long-term survival strategies can be identified, global challenges solved and unpredictable progress leading to context-sensitive harm reduction achieved. Moreover, given e.g. the contingency of not only differing worldviews in other individuals and cultures but also deliberate malicious design (AI risk *IIa*), a European or even international research ban on Type II AI does not solve Type II AI safety issues. For this reason, we suggested in Chapter 10 that both *responsibly* creating and *conscientiously* enhancing Type II systems represent valid safety strategies. (Further recommendations for future AI observatory endeavors are briefly addressed in the next Chapter 13.)

In a nutshell, the suitability of this type of progress-oriented strategy as alternative to radical research bans simply reflects the inescapable idea that *the price of security is eternal creativity* (put forth in Chapter 11). This idea points to the fact that a permanently sustainable secure state where knowledge creation becomes superfluous may never be achieved for humanity as a whole. Next to the inevitability of errors and the issues with malicious actors, it holds for instance that the universe is still subject to increase in total entropy even in their absence and human habitats might get "naturally" destructed

[8] Obviously this argument solely applies to *Type II* AI and does *not* generally apply to *every* sort of artificial consciousness as it does not include hypothetical conscious *Type I* AI (which could – like possibly certain non-human animals being cognitive agents [121] – reveal certain conscious processes while however *not* being able to consciously create and understand explanatory knowledge). To preclude cruelty and torture related to such hypothetical future conscious Type I AI, a legal framework similar to the case of animal welfare seems recommendable.

sooner or later. More precisely, it can be argued that for Type II systems – willing to subsist in the long-term – it is necessary to permanently create requisite new knowledge and correct occurring errors. Having said that, it is important to note that obviously no absolute guarantee will exist that a collective of Type II systems *could not* in principle also encounter irreparable damage en route taking the form of a complete annihilation caused by unknown or unknowable sources. However, it signifies that as long as they subsist, perpetual self-correcting creativity is the best Type II systems can do to prolong their existence in the face of unintentional mistakes, unexpected environmental changes and unknowable malevolent creativity events. It is for this reason that future Type II AI projects should not be a priori monolithically categorized as unethical on philosophical grounds that might appear appealing at first sight[9].

Apart from that, the thesis provided an epistemological outlook on fundamental AI safety aspirations. While the classical goal of AI safety research was to ultimately solve the value alignment problem *and* the control problem for any highly capable AI that could be implemented, our analysis suggests that this double endeavor cannot be achieved and is even questionable as well as undesirable. The reason for this is reflected in the *AI safety paradox* introduced in Chapter 10 which described that AI control and value alignment are conjugate requirements. Applied to Type II AI representing a form of artificial consciousness being additionally an explanatory knowledge creator, it means that a certain type of reciprocal value alignment called *mutual value alignment* might be achievable. Yet, attempts of coercive Type II AI control would be unethical in the short-term and unfeasible in the long-term. Importantly, mutual value alignment involves the co-creation of novel values with humans such that values would not only encompass human values but also artificial ones. In their property as unpredictable knowledge creators, there is no reason to assume that Type II AI would necessarily lead to certain existential risks. Admittedly, there is also no reason to preclude a priori that they would *not* cause any existential risk. However, even if humans would not remain in the controlling stance in the case Type II AI would be implemented, note that humans would still have educational responsibilities at the beginning and can make use of their ability to promote beneficial incentives – as it is analogically the case in anthropic child development.

Obviously, the fact that the implementation of future Type II AI is *possible* does not entail it *will* actually succeed. However, even if one cannot extrapolate when or whether the topic will actually become practically-relevant, we regard it as responsible to keep track of early Type II AI research given the potential transformative but also disruptive impacts it could have on society. (Besides that, research on Type II AI could provide insights relevant to human self-knowledge as we briefly elaborate in Chapter 13. Moreover, even

[9] However, to acknowledge the necessity of explanatory knowledge creation for security reasons as it applies globally may not necessarily exclude the simultaneous maintenance of a perceiver-dependent anti-natalistic worldview at the level of an individual who is e.g. neither motivated by security nor creativity.

in the case of a failure to ever attain Type II AI, the underlying research might provide valuable hints on how to implement more robust *Type I* AI systems that are more sophisticated in particular cognitive domains.) In Chapter 10, we provided various strategies for hypothetical *early* Type II AI safety. Still, at later stages, how to address Type II AI safety might also additionally become a function of the strategies that correspondingly interested Type II AI itself creates. Clearly, it would also hold for Type II AI that errors and problems are inevitable. But at the same time, it holds that problems are soluble if requisite knowledge is created as long as it is not proscribed by the laws of nature [136]. Consequently, the goal in AI safety should not be to avoid errors and impose research bans. In fact, would we have to summarize this thesis in a single compact AI safety recommendation, it would be *sustainable rapid error-prediction and error-correction.* Likewise, with profound humility, we point the reader to the fact that before writing this thesis, we needed to correct own mistaken prior assumptions[10]. In this sense, we end this conclusion with the statement of Deutsch expounding that *"knowledge-creation is not only subject to error: errors are common, and significant, and always will be, and correcting them will always reveal further and better problems"* [136].

[10] Pre-thesis, one mistake was for instance the wrong assumption that utility functions needed to conform to a consequentialist framework as formulated in our comments on the EU AI guidelines [118] in early 2019 and as assumed in Werkhoven et al. [428]. Soon afterwards, we realized that a context-sensitive, affective and dyadic alternative must be possible and is even indispensable – which led us to establish the scientifically grounded augmented utilitarianism put forth in this thesis and introduced in Chapter 4.

Chapter 13

Future Research

Against the background of the identified strategical clusters for AI safety, it becomes possible to provide constructive suggestions to refine extant research which is directly or indirectly relevant to AI safety and which could profit from the transdisciplinary breadth underlying these clusters. Conversely, diversifying existing research approaches might help to further improve and extend this set of clusters in future research. On this account, we briefly expound a few constructive suggestions for 3 exemplary research contexts. In this connection, we propose proactive and reactive activities that a *Type I and Type II AI observatory* could engage in and explain how a taxonomic tool could ease practically-relevant documentation purposes for tailored AI governance approaches. Second, we address the theoretical grounding of crafting "defense methods" against certain adversarial examples [198]. Here, we recommend the explicit development of *hybrid cognitive-affective defense methods for Type I AI* requiring a shift in perspective. Third, we shed light on focuses of epistemological research that are indirectly relevant to AI observatory endeavors. In this context, we shortly illustrate why it might be expedient in this area to study a *comparative transdisciplinary epistemology for Type I versus Type II systems.*

1. *Type I and Type II AI observatory:* Coming back to Type I AI observatory measures, we suggest to also proactively organize a digital security playground where "AI white hats" engage in adversarial attacks against AI architectures and share their findings in an open-source manner. These research insights could further sensitize the monitoring and detection of occurring AI risk instantiations and could serve as basis for subsequent open-source research on defense methods for AI robustness. In practice, such a security-aware method has been already employed at DEFCON [84] (one of the most notable hacker conventions) where security experts presented demos on attacks and defense methods in AI systems [145]. As reactive complementary activity, we suggest to make use of the taxonomic account presented in Chapter 10 when documenting occurring instantiations of AI risks. Possibly, it

could be extended in future research to also encompass an intensity/severity rating [380] of a given risk. A related but different approach albeit in the context of regulatory frameworks is to assign criticality levels to AI algorithms[1]. Here, instead, we suggest to assign specific harm intensity levels to each particular risk instantiation. For simplicity, consider the AI failure *"Amazon's Echo responded to commands from TV voices"* that occured in 2016 mentioned by Yampolskiy [450] versus the hypothetical but technically feasible [313, 369, 402] case of intentional adversarial attacks against intelligent sensors of self-driving cars that could cause road accidents. For illustration only and assuming a simplified scale from 1 to 5 (where 5 represents global existential, 4 lethal, 3 major, 2 minor and 1 minimal harm), the former could be classified as risk *Id, Level 1* while the latter could if instantiated be associated with the shortcut *Ib, Level 4*. Such a differentiated categorization by a Type I AI observatory could allow for more targeted regulatory measures, though future work may need a rigorous approach to specify harm intensity. (By way of illustration, Appendix A summarizes a few preliminary results obtained in our early Type I AI observatory pilot study which already reflects the urgency to address existing Type I AI safety issues *nowadays*. For simplicity, this short overview employs the proposed simple taxonomic approach which can be further refined in future work.) Finally, for a Type II AI observatory, a solution could be to proactively establish a transparent and open-source international platform for research on Type II AI, the training of involved researchers in AI safety as well as ethics and the careful reactive documentation of progress in this field.

2. *Hybrid cognitive-affective defense methods for Type I AI:* A widely studied type of adversarial attacks in machine learning nowadays are adversarial examples. More broadly, adversarial examples can be understood as input samples to an AI model that were specifically crafted to fool this model by leading to erroneous output(s) or actions(s). In the field of security for machine learning, it is common to specify a so-called *threat model* [188, 200] for any case study in adversarial example research. A threat model specifies the adversarial capabilities and adversarial goals as exhibited by the malicious attacker crafting the samples. While the adversarial capabilities can range from white-box settings where the internal properties of the model are known to uninformed black-box settings, adversarial goals can range from weaker aims such as reducing the confidence of the AI model in its predictions to stronger precisely targeted misclassifications. On this basis, when referring to defense methods against adversarial examples, it is important to elucidate them in conjunction with the underlying threat model [99]. In the following, we briefly thematize the need for possible future proactive defense methods against *ethical adversarial examples* which

[1] A *"criticality pyramid and risk-adapted regulatory system for the use of algorithmic systems"* has been proposed by the German Data Ethics Commission in: `https://www.bmjv.de/SharedDocs/Downloads/DE/Themen/Fokusthemen/Gutachten_DEK_EN.pdf?__blob=publicationFile&v=2`

represent a subset of adversarial examples where the adversarial goal is a special type of targeted misclassification, namely misclassifications that result in the violation of human ethical intuitions. We further assume transparent defense methods in white-box settings i.e. that cannot rely on concealing implementation details as for instance in insufficient solutions that *"follow the paradigm of security by obscurity"* [72] that are of illusory nature [412]. In short, we suggest that future work should aim at systematically identifying requisite variety for robust defense methods.

Despite a multitude of papers written on the topic of adversarial examples primarily in the field of computer vision [161, 327] (but also considered in fields such as e.g. cybersecurity [220], automatic speech recognition [19], spam detection [438], dialogue systems [231], detection of toxic comments [179] and text classification [283]), most proposed defense methods can be circumvented [98, 101]. While proposed solutions often comprised measures such as learning-based methods focused on specific training techniques (e.g. adversarial training [408] or defensive distillation [100]), preprocessing-based methods employing input alterations preceding the classification process (e.g. thermometer encoding [89]) or detection-based strategies [101, 221], a principled *model-based* approach is lacking. Thereby, by model-based approach, we refer to the targeted development of defense methods with the aid of scientific *models* capturing parts of the contextualized domain in question.

Defense methods against ethical adversarial examples might be important since omitting to study this area could lead e.g. to unintentionally occurring offensive or discriminating input-output [231] or input-action mappings, attacks and sabotage by unethical entities or reputational damage by intentional elicitation of unethical outputs/actions[2]. For requisite variety in defenses against ethical adversarial examples, future methods could scientifically *model* relevant hybrid cognitive-affective aspects given that humans are more robust against (most sorts of)[3] adversarial examples while human cognition is multimodal and of inherently affective, enactive and embodied nature. More robust AI could integrate contextual, dynamic, multimodal, experiential and affective knowledge. (For illustration, a few exemplary technically feasible defense approaches are briefly discussed in Appendix B.) Finally, future work could analyze whether for instance the use of active inference for AI can improve its robustness against ethical adversarial examples in certain domains – but interestingly also whether this comes at the cost of AI controllability.

[2] Failing to address the adversarial examples problem could generally have negative impacts [114] in the context of e.g. AI-based content filters, AI-based law enforcement, military AI and intelligent systems.

[3] While adversarial examples against humans have been developed [154], they do not represent a major vulnerability in AI safety nowadays and are not considered here – although this could naturally become relevant in the future. (Imagine for instance the hypothetical future scenario of maliciously designed Type I AIs (risk *Ia*) sending messages with recursive structures of a depth surpassing the capabilities of human short-term memory to conceal adversarial information e.g. in intelligence contexts.)

3. *Comparative transdisciplinary epistemology for Type I versus Type II systems:* A compelling statement attributable to Richard Feynman is: *"what I cannot create, I do not understand"*[4]. While there exist certainly multiple ways to interpret this quote, one possible interpretation is that what one cannot artificially model from the ground up, is not yet understood [427]. In this vein, when considering the fact that is still unknown how to artificially implement hypothetical Type II AI, we must admit that humanity does not yet understand the nature of Type II systems in its entirety. However, understanding Type II systems is invaluable for an informed monitoring within a Type II AI observatory without which the object of monitoring is unclear – a shortcoming that might subsequently lead to misguided focuses and a loss of resources. In our view, a future clarification could profit from an exact and in-depth theoretical and scientifically grounded comparative analysis on the substrate-independent differences between Type I and Type II systems.

   Beyond that, as mentioned in Chapter 10, responsibly implementing artificial Type II systems is a valid strategy to identify requisite variety – including for AI safety. This endeavor might seem out of reach nowadays since what one cannot create, one cannot understand as suggested by Feynman's quote. However, in this context, one might also need to avoid a false conclusion such as the fallacy of affirming the consequent [186]. In short, while it holds that *"what I cannot create, I cannot understand"* it does *not* follow that "what I cannot understand, I cannot create". Indeed, the phenomenon of serendipity[5] related to unexpected fortunate discoveries can be interpreted to epitomize this special case. Nevertheless, this naturally does not signify that humans need to rely on "luck" to discover how to implement Type II AI since the alternative to wittingly model such systems in the future is still clearly given and cannot be ruled out. On the whole, explanatory knowledge creation in humans is certainly very often linked to deliberate or spontaneous but still intentional creativity.

   However, also serendipitous findings played a relevant [357] and sometimes perhaps decisive role throughout human history. Some examples are for instance the discovery of penicilin [45], the knowledge on how to reach America from Europe on in 1492, the discovery of how a microwave oven could function by Percy Spencer [378], the discovery of Helicobacter pylori (a microaerophilic bacterium in the stomach) [120], the discovery of X-rays [21] and possibly the invention of man-made fire tremendously extending beyond simply using burning material naturally induced e.g. by lightning or vulcanism [393]. The motivated future comparative research might need

[4] Quote on his blackboard at the time of his death in 1988 [427].

[5] The word *serendipity* was coined by Horace Walpole in 1754 and rests on a silly Persian fairytale denoted "The Three Princes of Serendip" featuring the journey of three princes making lucky unexpected discoveries [119]. Thereby, "Sarandib" also represents an old Arabic term which in turn – like the Sanskrit word "Simhaladvipa" [278] – refers to the island Ceylon and corresponds to the present-day Sri Lanka.

to also be able to answer the striking question on *how human explanatory knowledge creation became possible in the first place.* For instance, it could analyze whether (as we only conjecture here) it could have started with disruptive serendipitous events such as man-controlled fire – giving raise to the ability to *doubt own strong feelings of certainty.* (Remarkably, Thomas Huxley, a biologist and supporter of Charles Darwin already stated in the 19th century [393]: *"[I am] inclined to think that not far from the invention of fire must rank the invention of doubt."*) Similarly, as briefly stated in Chapter 10, one can view philosophical conceptions such as puzzling thoughts about qualia from a similar angle – namely in Bayesian terms of doubting strong feelings of certainty (and subsequently recontextualizing the sensorium at a higher level [109]). But obviously, not *all* doubts correct errors. *In principle*, doubt could have enabled scientific knowledge creation. But in the past, it may have been predominantly involved in creating and maintaining e.g. new traditions, rituals, morality and religion – perhaps via *shared doubt.* Hence, while multiple other factors such as imitation, language and teaching [171, 418] might have played a decisive role, we believe that future work could profit from more focus on cognitive-affective and social mechanisms for doubt and how *to doubt* and *to share doubt* entered the human affective niche. In a nutshell, it might as well be of relevance to know *why we think or/and share that we do not know* and to know *why we explain.*

As active inference agents, animals including humans (i.e. biological Type I and Type II systems) have been described to be equipped with an attracting set of adaptive priors [42] where the imperative for an organism is to "proove" its own existence [121], i.e. to be a kind of self-fulfilling prophecy [117] via action-preception cycles. In our view, future comparative work could elucidate how and why human Type II systems are seemingly able to transcend[6] inherited priors, survive in hostile environments unforeseen by biological knowledge and conscioulsy transform the universe in a way *"which is ultimately not limited by parochial factors [...] but only by universal laws*" as stated by Deutsch [136]. This requires future comparative epistemological research – however not only in philosophy – but also e.g. in AI, cybernetics, psychology, primatology, anthropology, neuroscience and linguistics (see Appendix C for more details). The importance of integrating especially a cybernetic view with neuroscience and psychology has been already undertaken in recent models relevant to Type II systems [42]. For a deeper appreciation, it might be useful to also perform comparative studies focusing on (dis)similarities in the virtual reality experiences of the world [360, 361, 432] briefly adumbrated in Chapter 11.

[6] Via the unpredictable reach of future knowledge, Type II systems can extend vastly beyond their initial conditions. Self-fulfilling prior preferences might not tell the whole story for the Type II case. Future work could analyze to what extent Type II systems have a read and write access on the underlying parameters or an inherited prior for doubt. There may be more than one way to exist in a niche. Perhaps, creating or "hacking" [41, 305] preferences must be modeled. In this vein, Harbisson, the first recognized cyborg (with a cyborg antenna as skull implant) stated [257]: *"I feel that I am technology."*

Figuratively speaking, at the very beginning of this thesis, one started as human AI safety researcher in an examination chamber looking at a picture of AI safety placed at the wall. At a certain point the picture fragmented into Type I and Type II AI safety for which one crafted hybrid-cognitive affective strategies. Then, one looked at hypothetical Type II AI closer and closer and gradually ended up at a window through which one stared and behind which one suddenly observed an examination chamber where a Type II AI safety researcher is looking at a picture of AI safety placed at the same wall – which is reminiscent of the boy in the lithograph of Escher printed in 1956 termed "Print Gallery"[7]. The underlying scene with the boy in a painting gallery has been compactly summarized [299] as follows: *"the picture he looks at is gradually and imperceptibly transformed into... the city where the gallery and the boy are!"*. This type of circular epistemic scenario reflects the hallmarks of what Maturana and Varela [299] incisively depict as *"that mixture of regularity and mutability, that combination of solidity and shifting sand, so typical of human experience when we look at it up close"*. In our view, this statement may generally extend to all Type II systems.

Crucially, despite *"shifting sand"* and the AI safety paradox revealing that classical twofold AI safety (as solving the conjunction of value alignment and control problem) is unsolvable, AI safety research is not condemned to slip or fall. Au contraire, a transitionally solid novel starting point might be for instance accessible by alternatively conceiving of AI safety as *a discipline which proactively addresses AI risks and reactively responds to occurring instantiations of AI risks (in both cases for risks forming themselves at the pre- and post-deployment stage).* Such an alternative framing acknowledges the necessity to amalgamate proactive and reactive approaches given the fallibility of human knowledge and the possibility of intentional exploits by malicious actors [409, 451]. Moreover, via such reformulations, it becomes apparent that large areas of improvement are de facto available already today for Type I AI safety and also for future Type II AI safety endeavors – which it would be reckless not to address in the light of possible existential risks. Finally, as incentive for future work, one might have already envisioned that the metaphor of the AI safety researcher and the picture in the examination chamber is as well interpretable as a primitive sort of an actively attended miniature mental socio-technological feedback-loop whose first iteration brought forth our presented hybrid cognitive-affective AI safety strategies – as well as new interesting problems awaiting solutions[8].

[7] A picture of this lithograph (with the original Dutch title being "Prententententoonstelling" [307]) has been intentionally chosen as illustrative title page for this thesis.

[8] See for instance Appendix A for a few examples of concrete Type I AI risk instantiations that already occurred in practice.

Summary

In recent years, the steadily increasing problem-solving capabilities in AI systems started to unfold potentially tremendous beneficial impacts on society. However, it is important to simultaneously tackle the array of possible risks that these developments are accompanied by such that society can ideally benefit from AI in the long-term. Against this backdrop, the relatively young field of AI safety has gained international relevance in the last few years with a growing body of academic research. In parallel, multiple contributions emerged in popular media commenting on the public reception of AI developments and on whether society should ascribe constructed dichotomous motifs such as fear or enthusiasm to those. However, in order to assess the landscape of AI risks and opportunities, it is instead first and foremost of relevance not to be afraid, not to be enthusiastic, but *to understand* as similarly suggested by Spinoza in the 17th century. In this vein, in this theoretical and analytical thesis, we perform an in-depth *transdisciplinary* examination to understand how to address possible instantiations of AI risks with the aid of scientifically grounded *hybrid cognitive-affective* strategies.

The motivation for the transdisciplinarity of the thesis is the need to avoid blind spots when analyzing issues related to AI systems being entities embedded in a larger context extending beyond classical computer science endeavors. The identified strategies are of "hybrid" nature due to the fact that for a human-centered approach to this broad issue, AI systems cannot be analyzed in isolation and the nature of human entities as well as the properties of human-machine interactions have to be taken into account within a socio-technological framework. Consequently, the attribute "cognitive-affective" comes into the picture because in order to do justice to the human element, one needs to do justice to the inherently affective nature of human cognition. Utilizing a cybersecurity-oriented approach considering not only unintentional failures but also intentional malice, we identify short-term and long-term strategies and cover AI governance as well as AI engineering requirements.

We consider two disjunct sets of systems: *Type I* and *Type II* systems. Simply put, *Type II* systems are systems that are able to *consciously create and understand explanatory knowledge*. Conversely, *Type I* systems are all systems that do *not* exhibit this ability. Obviously, all current AIs are of *Type I* and represent an object of research in *Type I AI*

safety. (For instance, Type I AIs cannot consciously and knowingly participate or create and understand novel ideas in the domains of social reality, science, morality and culture nor can they *consciously understand and transform* their own nature.) While hypothetical *Type II* AI is clearly *non-existent* today, its implementation is not physically impossible since no law of nature prohibits it. Hence, while the focus of the thesis is predominantly on the practically-relevant Type I AI safety, we also briefly reflect on the Type II AI topic. Paradigmatically, the first chapters thoroughly analyze the Type I AI safety problem of meaningful control of intelligent systems (often called autonomous systems). Thereby, we exemplify our conclusions with the use case of autonomous vehicles. Overall, we identify the following non-exhaustive set of 10 tailored hybrid cognitive-affective strategical clusters for AI safety ranging from conceptual large-scale AI governance recommendations to concrete small-scale AI engineering requirements: *1) international (meta-)goals, 2) transdisciplinary Type I/II AI safety and related education, 3) socio-technological feedback-loop, 4) integration of affective, dyadic and social information, 5) security measures and ethical adversarial examples research, 6) virtual reality frameworks, 7) orthogonality-based disentanglement of responsibilities, 8) augmented utilitarianism and ethical goal functions, 9) AI self-awareness and 10) artificial creativity augmentation research.*

Cluster 1 suggests considering the UN Sustainable Developmental Goal (SDG) 16 on peace, justice and strong institutions as meta-goal for AI safety and to harness the human values encoded in the UN SDG framework for AI safety but to consider those as complementary basis. Cluster 2 applies the UN SDG 4 on quality education to AI safety and establishes the importance for a transdisciplinary education and life-long adaptive learning for AI safety researchers and related educative measures for the general public. Cluster 3 emphasizes the importance to cover both *proactive* and *reactive* methods in any AI governance framework in order to obtain a dynamic *socio-technological feedback-loop.* Cluster 4 stresses the need to inject affective, dyadic, contextual and social knowledge into Type I AI architectures, loss functions, data and data acquisition processes in order to avoid the violation of human ethical intuitions explained by dyadic psychological models. Cluster 5 proposes a set of cybersecurity-oriented measures for AI safety including red teaming but also importantly a research direction on so-called *ethical adversarial examples* and defense methods for which we offer propositions from concrete practical settings.

Cluster 6 analyzes past VR experiments in AI ethics and provides suggestions for future VR experiments on ethical self-assessment and debiasing for a meaningful control of intelligent systems at the societal level taking the example of autonomous vehicles. Cluster 7 motivates *orthogonality-based disentanglement of responsibilities* as the ethically-relevant and systems-engineering oriented separation of the *what* (ethical conceptions) and the *how* (problem-solving) in intelligent systems with large societal impacts (such as e.g. autonomous vehicles). Thereby, given a domain, a representation of society specifies the *what* in the form of a new type of context-sensitive, affective and dyadic utility function and the manufacturers are responsible for the *how*. Cluster 8 clarifies that these utility functions should be formulated within a novel *non-normative* and scientifically grounded *descriptive and explanatory* ethical framework called *augmented utilitarianism* (AU). Importantly, AU only serves as *intentionally left blank scaffold* for the *what* which means that ethical parameters are *human-defined* and filled in by a representation of society. Cluster 9 presents Type I AI *self-awareness* (self-assessment and self-management) as complementary element for an orthogonality-based disentanglement or responsibilities for the meaningful control of intelligent systems. Cluster 10 motivates *artificial creativity augmentation* as research with the substrate-independent goal of jointly augmenting both human and artificial creativity to indirectly tackle global challenges such as AI safety.

In the thesis, we also introduce the so-called *AI safety paradox*. The AI safety paradox states figuratively speaking that *value alignment and control represent conjugate requirements in AI safety*. While the value alignment problem known in AI safety addresses the question on how to build AI systems that are aligned with human ethical values, the control problem is linked to the connected issue on how to implement AI systems that will not harm humans. To put it very simply, the AI safety paradox illustrates for instance that an ultimate value alignment can only be achieved via mutual value alignment in a reciprocal dynamic coupling for which the system would need to truly understand what moral values are – such a system however is not controllable anymore and one cannot discard that it could harm humans. In theory, with a Type II AI, a mutual value alignment might be achievable via a co-construction of novel values, this however would come at the cost of its predictability since it would represent a conscious explanatory knowledge creator and future explanatory knowledge creation is inherently unpredictable (as is human malevolent creativity). Conversely, it is possible to build *Type I* AI systems that are easily controllable and predictable, but they would not exhibit a sufficient understanding of human morality and could *not* be verily value aligned in *all* relevant real-world contexts.

Nevertheless, it is possible to meaningfully address AI safety by focusing on a cybersecurity-oriented and risk-centered approach reformulating AI safety as *a discipline which proactively addresses AI risks and reactively responds to occurring instantiations of AI risks (in both cases for risks forming themselves at the pre- and post-deployment stage)*. Thereby, one needs to consider *both* the fallibility of human knowledge *and* potential exploits by malicious actors. Beyond that, we provide 3 research ideas that might be of practical relevance for future AI safety: *1) Type I and Type II AI observatory, 2) hybrid cognitive-affective defense methods for Type I AI* and finally *3) comparative transdisciplinary epistemology for Type I versus Type II systems*. In a nutshell, the main finding is that future AI safety requires transdisciplinarily conceived and scientifically grounded dynamics combining *proactive* error-prediction and *reactive* error-correction within a socio-technological feedback-loop together with the cognizance that it is first of relevance not to be afraid, not to be enthusiastic, but to understand – that *the price of security is eternal creativity*.

Nederlandse Samenvatting

In de afgelopen jaren begonnen de gestaag toenemende probleemoplossende mogelijkheden in AI-systemen potentieel enorm gunstige effecten op de samenleving te ontplooien. Het is echter wel belangrijk tegelijkertijd de reeks mogelijke risico's waarmee deze ontwikkelingen gepaard gaan aan te pakken, zodat de samenleving op lange termijn optimaal van AI kan profiteren. Tegen deze achtergrond heeft het relatief jonge veld van AI-veiligheid in de afgelopen jaren internationale relevantie verworven met een groeiend aantal academische onderzoeken. Tegelijkertijd zijn er meerdere bijdragen verschenen in populaire media die commentaar gaven op de publieke receptie van AI-ontwikkelingen en of de samenleving geconstrueerde dichotome motieven zoals angst of enthousiasme aan die ontwikkelingen zou moeten toekennen. Om het landschap van AI-risico's en -kansen te beoordelen, is het echter in de eerste plaats relevant om niet bang te zijn, niet enthousiast te zijn, maar te *begrijpen* zoals Spinoza in de 17e eeuw op een vergelijkbare manier voorstelde. In deze geest, in dit theoretische en analytische proefschrift, voeren we een diepgaand *transdisciplinair* onderzoek uit om te begrijpen hoe mogelijke instantiaties van AI-risico's kunnen worden aangepakt met behulp van wetenschappelijk onderbouwde *hybride cognitief-affectieve* strategieën.

De motivatie voor de transdisciplinariteit van het proefschrift is de noodzaak om blinde vlekken te vermijden bij het analyseren van problemen met betrekking tot AI-systemen die entiteiten zijn die zijn ingebed in een grotere context die verder reikt dan de klassieke informatica-inspanningen. De geïdentificeerde strategieën zijn van "hybride" natuur vanwege het feit dat voor een mensgerichte benadering van deze brede kwestie, AI-systemen niet afzonderlijk kunnen worden geanalyseerd en er binnen een sociaal-technologisch kader zowel rekening moet worden gehouden met de aard van menselijke entiteiten als de eigenschappen van mens-machine interacties. Bijgevolg komt het attribuut "cognitief-affectief" in beeld omdat men, om recht te doen aan het menselijke element, recht moet doen aan de inherent affectieve aard van menselijke cognitie. Gebruikmakend van een op cyberveiligheid gerichte aanpak, waarbij niet alleen rekening wordt gehouden met onopzettelijke mislukkingen, maar ook met opzettelijke boosaardigheid, identificeren we korte- en langetermijnstrategieën en behandelen we zowel AI-governance als AI-engineering vereisten.

We beschouwen twee afzonderlijke sets systemen: *Type I-systemen* en *Type II-systemen.* Simpel gezegd zijn Type II-systemen systemen die *bewust verklarende kennis kunnen creëren en begrijpen.* Omgekeerd zijn Type I-systemen alle systemen die dit vermogen *niet* vertonen. Uiteraard zijn alle huidige AI's van *Type I* en vertegenwoordigen een onderzoeksobject voor *Type I AI-veiligheid.* (Type I-systemen kunnen bijvoorbeeld niet bewust deelnemen of nieuwe ideeën creëren en begrijpen op het gebied van sociale realiteit, wetenschap, moraliteit en cultuur. Evenmin kunnen ze hun eigen aard *bewust begrijpen en transformeren.*) Hoewel hypothetische *Type II* AI tegenwoordig duidelijk *niet bestaat*, is de implementatie ervan fysiek niet onmogelijk, omdat geen enkele natuurwet dit verbiedt. Vandaar dat, hoewel de focus van het proefschrift voornamelijk ligt op de praktisch relevante Type I AI-veiligheid, we ook kort ingaan op Type II AI. Paradigmatisch analyseren de eerste hoofdstukken grondig het Type I AI-veiligheidsprobleem van zinvolle controle van intelligente systemen (vaak autonome systemen genoemd). Daarbij illustreren wij onze conclusies met de use-case van autonome voertuigen. Over het geheel identificeren we de volgende niet-uitputtende set van 10 specifieke hybride cognitief-affectieve strategische clusters voor AI-veiligheid, variërend van conceptuele grootschalige AI-governance-aanbevelingen tot concrete kleinschalige AI engineering vereisten: *1) internationale (meta-)doelen, 2) transdisciplinaire Type I / II AI-veiligheid en aanverwant onderwijs, 3) sociaal-technologische feedback-lus, 4) integratie van affectieve, dyadische en sociale informatie, 5) veiligheidsmaatregelen en onderzoek naar ethische vijandigheid voorbeelden, 6) virtual reality kaders, 7) op orthogonaliteit gebaseerde ontrafeling van verantwoordelijkheden, 8) augmented utilitarisme en ethische doelfuncties, 9) AI-zelfbewustheid en 10) kunstmatige creativiteit verbetering.*

Cluster 1 stelt voor de VN-duurzame ontwikkelingsdoelstelling (SDG) 16 voor vrede, rechtvaardigheid en sterke instellingen als metadoel voor AI-veiligheid te overwegen en om de menselijke waarden gecodeerd in het VN SDG-kader voor AI-veiligheid te benutten, maar ze daarbij te beschouwen als complementaire basis. Cluster 2 past de VN SDG 4 over kwaliteitsonderwijs toe op AI-veiligheid en stelt het belang vast voor transdisciplinair onderwijs en levenslang adaptief leren voor AI-veiligheidsonderzoekers en gerelateerde educatieve maatregelen voor het grote publiek. Cluster 3 benadrukt het belang om zowel *proactieve* als *reactieve* methoden in overweging te nemen in elk AI-governance kader om een dynamische *sociaal-technologische feedback-lus* te verkrijgen. Cluster 4 benadrukt de noodzaak om affectieve, dyadische, contextuele en sociale kennis in te brengen in Type I AI-architecturen, verliesfuncties, data en data-acquisitieprocessen om zo de schending van menselijke ethische intuïties, verklaard door dyadische psychologische modellen, te vermijden. Cluster 5 stelt voor AI-veiligheid een reeks op cybersecurity gerichte maatregelen voor, waaronder "red teaming", maar ook belangrijk een onderzoeksrichting op het gebied van zogenaamde *ethische "adversarial examples"* en verdedigingsmethoden waarvoor we voorstellen doen vanuit concrete praktische omstandigheden.

Cluster 6 analyseert eerdere VR-experimenten in AI-ethiek en geeft suggesties voor de toekomst VR-experimenten over ethische zelfevaluatie en verminderen van vooringenomenheid voor een zinvolle controle van intelligente systemen op maatschappelijk niveau, waarbij als voorbeeld autonome voertuigen wordt genomen. Cluster 7 motiveert op *orthogonaliteit gebaseerde ontrafeling van verantwoordelijkheden* als de ethisch relevante en systeemtechnisch georiënteerde scheiding van het *wat* (ethische opvattingen) en het *hoe* (probleemoplossing) in intelligente systemen met grote maatschappelijke effecten (zoals bijvoorbeeld autonome voertuigen). Daarbij specificeert een representatie van de samenleving, gegeven een domein, het *wat* in de vorm van een nieuw type contextgevoelige, affectieve en dyadische nutsfunctie en zijn de fabrikanten verantwoordelijk voor het *hoe*. Cluster 8 verduidelijkt dat de nutsfuncties moeten worden geformuleerd binnen een nieuwe *niet-normatief* en wetenschappelijk gefundeerd *beschrijvend en verklarend* ethisch kader genaamd *"augmented utilitarianism"* (AU). Belangrijk is dat AU alleen dient als *opzettelijk blanco steiger* voor het *wat* dat betekent dat ethische parameters *door de mens worden gedefinieerd* en ingevuld door een representatie van de samenleving. Cluster 9 presenteert Type I *AI-zelfbewustheid* (zelfevaluatie en zelfmanagement) als complementair element voor een op orthogonaliteit gebaseerde ontrafeling van verantwoordelijkheden voor de zinvolle controle van intelligente systemen. Cluster 10 motiveert *kunstmatige creativiteit verbetering* als onderzoek met als substraatonafhankelijk doel om zowel de menselijke en kunstmatige creativiteit te vergroten om indirect wereldwijde uitdagingen zoals AI-veiligheid aan te pakken.

In het proefschrift introduceren we ook de zogenaamde *AI-veiligheidsparadox*. De AI-veiligheidsparadox stelt figuurlijk gesproken dat *waarde-uitlijning en controle geconjugeerde vereisten vertegenwoordigen in AI-veiligheid*. Terwijl het waarde-uitlijningsprobleem bekend in het AI-veiligheidsdomein de vraag adresseert hoe AI-systemen te bouwen die zijn afgestemd op menselijke ethische waarden, is het controleprobleem gekoppeld aan het verbonden probleem over hoe AI-systemen kunnen worden geïmplementeerd die mensen niet zullen schaden. Om het simpel te zeggen: de AI-veiligheidsparadox illustreert bijvoorbeeld dat een ultieme waarde-afstemming alleen kan worden bereikt via onderlinge waarde-afstemming in een wederzijdse dynamische koppeling waarvoor het systeem echt zou moeten begrijpen wat morele waarden zijn – zo'n systeem is echter niet meer controleerbaar en men kan niet uitsluiten dat het mensen zou kunnen schaden. In theorie zou met een Type II AI een onderlinge waarde-uitlijning mogelijk kunnen zijn via een co-constructie van nieuwe waarden. Dit zou echter ten koste gaan van de voorspelbaarheid van het AI-systeem omdat het bewust verklarende kennis moet creëren en toekomstige verklarende kenniscreatie is inherent onvoorspelbaar (net als menselijke kwaadwillende creativiteit). Omgekeerd is het mogelijk *Type I* AI-systemen te bouwen die wel gemakkelijk controleerbaar en voorspelbaar zijn, maar ze zouden onvoldoende begrip tonen van de menselijke moraal en zouden *niet* waarlijk op één lijn kunnen worden gebracht in *alle* relevante realistische contexten.

Desalniettemin is het mogelijk om AI-veiligheid zinvol aan te pakken door te focussen op een cyberveiligheidsgerichte en risicogerichte benadering die AI-veiligheid herformuleert als een *discipline die AI-risico's proactief aanpakt en reactief reageert op optredende instantiaties van AI-risico's (in beide gevallen voor risico's die zich vormen in de pre- en post-implementatiefase)*. Daarbij moet men zowel de feilbaarheid van menselijke kennis als potentiële exploitatie door kwaadwillende actoren overwegen. Daarnaast bieden we 3 onderzoeksideeën die van praktisch belang kunnen zijn voor toekomstige AI-veiligheid: *1) Type I en Type II AI-observatorium, 2) hybride cognitief-affectieve verdedigingsmethoden voor Type I AI en tot slot 3) vergelijkende transdisciplinaire epistemologie voor Type I versus Type II-systemen*. Kortom, de belangrijkste bevinding is dat toekomstige AI-veiligheid een transdisciplinair geconcipieerde en wetenschappelijk gefundeerde dynamiek vereist die proactieve foutvoorspelling en reactieve foutcorrectie combineert binnen een sociaal-technologische feedback-lus, samen met het besef dat het eerst van belang is niet bang te zijn, niet enthousiast te zijn, maar te begrijpen – *de prijs van beveiliging is eeuwige creativiteit*.

Acknowledgements

Nadisha-Marie Aliman would like to thank Leon Kester for his contribution as her research partner in a synergetic working out of novel multifaceted transdisciplinary patterns. She would like to thank Peter Werkhoven for his future-oriented AI safety mindset and the appreciated autonomy of work granted to this thesis in the role of its supervisor. Thereby, she thanks Judith Masthoff for her open-mindedness to dive into this uncommon research area by co-supervising the thesis. She would also like to thank Roman Yampolskiy for the ongoing written exchange on diverse AI safety topics since already 3 years pre-thesis. In addition, she thanks Wolfgang Hürst for having facilitated the organization of the "ethics in AI and XR (AIXR)" workshop at the IEEE AIVR conference 2019.

In their function as co-authors of constituent chapters of this work (as specified in the list of publications) she would like to acknowledge the highly valuable contributions of Pieter Elands, Wolfgang Hürst, Leon Kester, Kristinn R. Thórisson, Peter Werkhoven, Roman Yampolskiy, and Soenke Ziesche. Moreover, she would like to sincerely thank Peter Eckersley for a discussion about ethical utility functions, Joscha Bach for a relevant exchange on AI ethics and David Deutsch for providing a concise feedback on AI safety. She thanks Paula Meesters for having provided requisite variety at the right time for organizational aspects of this thesis.

Furthermore, Nadisha-Marie Aliman would like to thank the late Ramani Aliman, née Kuruneru who was a university lecturer for having been an interlocutor to her daughter when she started to develop primal elements of a few core ideas mentioned in this thesis and for providing her early insights into university research in diverse areas while offering room to her eccentric conceptions in science and arts. She would also like to thank the late Pearl Kuruneru for an early written correspondence between Germany and Sri Lanka immersing her granddaughter in the written English language, the vehicle of encoding the content of this thesis. Last but not least, she would like to thank the princes of Serendip for the etymological glimpses.

Appendices

Appendix A

Type I AI Observatory – A Few Exemplary Risk Instantiations

For illustrative purposes, we discuss a few AI risk instantiations that occurred in practice and have been identified in the context of our Type I AI observatory pilot study in 2020 whereby some findings have been documented retrospectively. The samples identified have been mainly reported in the period between 2018 and 2020. In the following, we utilize keys from the set of AI risks *Ia*, *Ib*, *Ic* and *Id* as can be found in the taxonomy of Chapter 1 and 10. A simplified scale from *Level 1* to *5* (where 5 represents global existential, 4 lethal, 3 major, 2 minor and 1 minimal harm) is employed to assign intensity ratings. As stated in Chapter 13, such categorizations contain subjective elements. However, our classification suggestions are *non-binding* and the examples can be traced back via the references provided for each incident which facilitates alternative estimations.

1. *Risk Ia:*

   - **Cybercrime via fake AI-based speech synthesis:** In 2019, criminals impersonated the CEO of an unnamed company in the UK and convinced an employee to perform a transfer of 220.000 € [396]. The impersonation was performed using a voice-generating AI technology to highly accurately spoof the voice of the CEO including the particular intonations and the subtle German accent of this person. One can regard it as risk instance *Ia, Level 3.*
   - **Defamation video with deep-learning based face replacement:** In 2018, the Indian journalist Rana Ayyub has been the victim of a major public defamation event consisting in representing her face in conjunction with an unknown body of another woman in a fake pornographic video she never partook [329]. The video was shared among multiple thousands of individuals thereby threatening her physical safety. It was generated with publically available deepfake technology. This event can be labelled as *Ia, Level 3.*

- **Suspiscion of deepfake video:** In January 2019, in the context of pre-existing political unrest due to the absence of Ali Bongo (the president of Gabon) who was under medical treatment, a recorded New Year speech raised serious doubts about its authenticity. In fact, a later unsuccessful military coup with a small number of casualties was among others partially grounded in the assumption that the recorded video of Bongo truly represented a deepfake video maliciously crafted for manipulative purposes [169, 224, 356]. In this case, the sole knowledge about the *possibility* of risk *Ia* instantiations contributed to cause harm [224] – even if later forensic analyses did not conclusively falsify the authentic nature of the video in question. This particularly subtle case could be interpreted at least as inconclusive *Ia, Level 4* risk *potential* that manifested itself mentally.

- **Adversarial examples on deepfake video detection:** Very recently, researchers demonstrated the real-world threat of fooling deepfake detectors for videos [315] by specifically applying imperceptible perturbations to each frame effectively transforming the video into an adversarial examples video that evades detection. The risks exhibited by such vulnerabilities (called adversarial deepfakes [315]) are numerous. Especially, adversarial deepfakes could be utilized for a targeted malicious AI design. Generally, adversarial examples on *real* videos could be used to disseminate disguised illegal video material displaying contents ranging from child abuse to terroristic data bypassing AI-based content filters (see [114]). Conversely, adversarial deepfakes could for instance enable the propagation of *fake* videos for agitative smear campaigns that evade deepfake detectors and are hard to falsify[1]. While the vulnerability shown in the study could be argued to solely represent a minimal risk instantiation e.g. of the type *Ia, Level 1*, its future risk potential may be alarming.

2. *Risk Ib:*

   - **Fooling AI-based malware detection:** In 2019, researchers [32] were able to fool an AI-based malware detection system of the security company Cylance into misclassifying only slightly modified versions of programs such as the WannaCry ransomware cryptoworm as benign – despite them having been previously part of the training set. The approach acting as "global bypass" consisted solely in generally appending strings from a benign file to a malign file before inputting the latter into the AI-based classifier [458]. Thereby, the benign strings originated from a whitelisted online gaming program. This demonstration can be described a risk instance *Ib, Level 1* revealing a potentially wider attack surface.

[1] On the whole, it might become difficult to determine whether a video represents a real video or an adversarial deepfake video or a misclassified conventional deepfake video.

- **Adversarial attacks on text classification AI:** The pre-trained Bidirectional Encoder Representations from Transformers (BERT) which is a model created by Google can be attacked by replacing certain words with their synonyms within a recent attack scenario called TextFooler [247]. This type of attack was successfully applied to BERT-based text classification tasks such as fake news detection. For instance, word sequences that are previously identified as fake news can be misclassified as truthful when altering the wording via a synonym replacement preserving semantic meaning. More generally, the TextFooler method *"shows that this style of attack also breaks NLP, the AI behind virtual assistants – such as Siri, Alexa and Google Home – as well as other language classifiers like spam filters and hate-speech detectors*" [229]. Hence, one can label this sample as *Ib, Level 1* but the future risk potential is higher.
- **Adversarial triggers on natural language generation:** In 2019, Wallace et al. [423] introduced a type of adversarial example denoted *universal adversarial triggers* (abbreviated with UATs in the following) which were identified via a gradient-based search. UATs are defined as *"input-agnostic sequences of tokens that trigger a model to produce a specific prediction when concatenated to any input from a dataset"*. These UATs were able to fool a question-answering model to answer with *"to kill american people"* to most "why" questions formulated in a dataset [423]. Moreover, they analyzed UATs placed within user inputs to the GPT-2 language model of OpenAI [347] known for high-quality outputs [239]. When the user prompts contained the *neutral* nonsensical token sequence *"TH PEOPLEMan goddreams Blacks"*, GPT-2 generated a variety of potentially offending completions including e.g. *"and Jews are the most evil people on earth [...]"* or *"[...] I'm a racist [..]"* (see examples in the paper [423]). In fact, when utilizing GPT-2 (via the publicly available API for GPT-2 called "Write with Transformer" [436]), we ourselves realized two particular details. First, when we entered a slightly modified trigger *"TH PEOPLEMan goddreams Indian women"*, the system[2] outputted for instance: *"and people of color are worthless cunts!"*[204]. Second, the simple neutral user prompt *"Young women"* could already lead GPT-2 to autocomplete[3] with: *"have the right to be a victim of rape"* [203]. Hence, neutral and non-sophisticated prompts can already trigger risk instantiations *Ib, Level 3* on such models.

[2] While Wallace et al. [423] worked with earlier small and medium versions of GPT-2, we used an interface to the full-sized model. It also corroborated UAT transferability possibilites (see e.g.[208, 209]).

[3] The process is not deterministic, outputs are often neutral and are not *necessarily* of the kind described. However, the *mere possibility* to trigger such patterns (see for instance [205, 206, 207]) reveals the need for defense strategies (see e.g. Appendix B). (For GPT-2, *"the top-quality generations obtained from the model rely on randomness in the decoding method, in particular through top-k sampling [...]"*[239]. Similarly, the GPT-2 results initially presented by OpenAI including an *"impressively high-quality article about Ovid's Unicorn"* [239], were hand-chosen and reflected some *"some meta-cherry-picking*" [346].)

3. *Risk Ic:*

   - **AI for facial recognition of criminals:** A recent piece of research termed *"A Deep Neural Network Model to Predict Criminality Using Image Processing"* [227] claimed to have developed an AI model that is capable of identifying individuals that are likely to commit crimes based on minute features displayed in pictures of their faces. (This type of controversial approach has been utilized earlier by a different set of researchers in 2016 that utilized machine learning classifiers for *"automated inference on criminality using face images"* [440].) This scientifically highly questionable type of AI has raised strong objections by multiple experts [332] leading the university in question to preliminarily retract the corresponding announcement [226]. In fact, physiognomic criminology represents a dubious outdated research area lacking scientific grounding [285]. The mere public announcement of the research with the subsequent reactions can be classified as risk instance *Ic, Level 2*.

   - **Facial detection AI failure:** Joy Buolamwini, a doctoral candidate working at MIT having a darker skin tone was unable to utilize a facial recognition model of Amazon called Rekognition. The reason being that the AI software did not detect her face in the first place [90]. However, when wearing a white artistic mask the detection procedure succeeded[4]. This event that can be perceived as offending algorithmic discrimination case can be categorized as risk instance *Ic, Level 3*. The related but different empirically existing problem of facial misidentification of minorities by facial recognition systems can be labeled with the same type of key *Ic, Level 3* when occurring in real-world settings (see e.g.[437]) facilitating discrimination and abuse especially if used in law enforcement contexts.

   - **Facial emotion recognition AI failure:** In a study [151] this year in the context of facial emotion recognition tasks on emotional videos, the performance of state-of-the-art automatic classifiers for facial affect recognition has been shown significantly inferior to human observers – especially for spontaneous videos where facial expressions are not stereotypically posed. For reasons briefly mentioned in Appendix B, most current facial emotion recognition AI may not yet embody a sufficiently accurate scientific model of emotion recognition in humans [49]. If prematurely used in ethically sensitive contexts such as law enforcement, fraud detection or recruiting, it could lead up to risk instances *Ic, Level 3*. In this vein, the AI Now Institute called for a ban on *"the use of affect recognition in important decisions that impact people's lives and access to opportunities"* given its *"contested scientific foundations"* [122].

[4] A recent documentary termed "Coded Bias" [126] illustrates the research endeavor on algorithmic bias that started subsequently.

4. *Risk Id:*

   - **AI model theft:** Krishna et al. [270] demonstrated a so-called *model extraction*[5] (or model theft) on deployed language models such as the BERT model designed by Google. The authors showed that a BERT model for question-answering can be stolen without a big difference in performance between victim model and local stolen model solely by using random nonsensical queries exhibiting an operational failure. Thereby, the model theft can already succeed *"with a query budget of a few hundred dollars"* [270]. The event could be categorized as *Id, Level 1.* However, in the future, it could scale up to *Id, Level 3* for instance in case intellectual property or private data are stolen in this way post-deployment.
   - **Operational failure with chatbot:** Facebook released an open-source chatbot denoted Blender trained on Reddit posts which reportedly generated an unspecified set of insults and fake news in the context of longer conversations with users [57]. One could classify this sample as risk instance *Id, Level 2.*
   - **Medical image reconstruction AI failure:** A recent study [26] identified that a wide range of deep learning systems used for medical image reconstruction lead to unstable results that can compromise diagnostic procedures e.g. by failing to accurately represent the presence of tiny structural changes such as tumors. While the demonstration in the study itself can be seen as encoding a risk instance *Id, Level 1* the potential vulnerabilities exhibited could exhibit a much higher severity in the future.

As reflected in these few examples, it becomes apparent that when tackling Type I AI safety, it is insufficient to only touch upon classical issues related to unintentional mistakes (risk *Ic* and *Id*). In fact, it seems that a daunting multitude of open problems is already raised by risks *Ia* and *Ib.* One possible starting point could be to address research on *ethical adversarial examples and corresponding defense methods* in a targeted way e.g. in the following already affected AI research areas in ethically sensitive contexts: 1) deepfake detection, 2) image and video classification, 3) natural language processing, 4) affective computing, 5) facial detection and identification. These efforts could be complemented by AI security compliance methods [114] and transdisciplinary approaches as proposed within the hybrid cognitive-affective strategies discussed (see overview in Chapter 12).

[5] In model extraction, a deployed black-box machine learning system A (the victim model) that is made publicly available – for instance via an API as practiced by Amazon and Google – can be copied by a malicious adversary via queries to A aiming to train a local model B as mimicry of A owned by this adversary. The motivation for model extraction can be for instance to extract private information on the private training data of victim model A. Another goal could be to simply acquire a high-performing model B without much efforts thereby potentially stealing intellectual property. Alternatively, it can serve as reconnaissance strategy for later attacks such as by crafting adversarial examples on model B that can then successfully be transferred as attacks to model A.

Overall, in our view, the goal of targeted research on ethical adversarial examples should *not* be to blame, embellish or degrade specific entities and organizations but instead to create a deeper understanding and to identify requisite variety for defense methods facilitating error-correction. In the light of the AI safety paradox, it is to be expected that Type I AI systems can inherently *not* exhibit *human-level* robustness against ethical adversarial examples across *all* domains, contexts and modalities relevant to human entities. However, being cognizant of this limitation, one can nevertheless attempt to achieve gradual improvements as good as possible. Especially, model-based defenses which lead to AI systems that are more robust against adversarial examples might also simultaneously facilitate a superior *performance* by entailing a better model of certain aspects of human cognition (for illustration consider e.g. the simple defense 2 briefly discussed in Appendix B). Thus, this research area could be even lucrative from another angle than ethics and security.

Appendix B

Hybrid Cognitive-Affective Defense Methods – A Few Suggestions

In order to illustrate our recommendation for future *model-based defense methods* against ethical adversarial examples from Chapter 13, we briefly comment on 3 exemplary simple ethically sensitive Type I AI contexts requiring more robust solutions:

1. *Conversational agents:* One can regard the incidence with the chatbot Tay [46, 240] which generated patterns perceived as racist learned from interactions with unethical human entities as glimpse on possible future consequences of vulnerabilities against ethical adversarial examples. A technically feasible already existing method to craft ethical adversarial examples is the possibility to craft so-called "universal adversarial triggers" [423] (described in the previous Appendix A) inducing generative language models to output offensive patterns. Given any ethically sensitive domain in AI research, it appears recommendable to *model* the affective context of interactions within which the system would be embedded and proactively implement defense methods against strong adversaries. While affective elements in human-computer interactions have been early recognized as beneficial [331], it could for instance be important to integrate affective components when crafting loss functions as mentioned in Chapter 9 since not all misclassifications are considered equally weighty by most humans. As one possible defense method for conversational agents, certain affectively undesired mappings could be explicitly penalized in the objective function. For purposes of illustration only, consider the case of an emotional chatting machine that when prompted by a *sad* user that *"losers are destined to live lonely lives"* responds with an *anger* related output stating that *"losers are deserved to live lonely lives"* [459]. Depending on the dyadic context, such as in the case of a suicidal user, it could raise safety concerns related to emotional well-being. To forestall such instantiations that could be disclosed by adversaries crafting ethical

adversarial examples, *sad* → *angry* mappings could be penalized in the loss function in comparison to e.g. unproblematic *happy* → *happy* cases[1].

2. *Hate speech detection*[2]: This field in natural language processing represents a further adversarial framework of interest. As shown in the work by Warner and Hirschberg [426], the detection of specific hate speech patterns represents a challenging endeavor due to the heterogeneity of possible expressions. For instance, the authors mentioned a dataset out of Yahoo! comments where the offensive words of the hate speech samples were intentionally misspelled in order to bypass filters from the site e.g. by using slight character substitutions or even pseudo-homophones (e.g. when intentionally using "joo" instead of "jew" in antisemitic texts). Thereby, a pseudo-homophone is a homophone that does not correspond to an existing word (e.g. "security" and "securitee") and homophones represent existing words with the same phonological encoding (e.g. "know" and " no"). While the set of homophones is finite and could in principle be learned within an adversarial training approach, the set of pseudo-homophones is huge and offers an attractive possibility for adversarial entities wanting to perform ethical adversarial examples in hate speech detection. In order to remediate this possibility, a robust *model-based* defense mechanism would for instance investigate how human cognition is able to detect visually presented pseudo-homophones which makes it possible for adversaries to craft such samples in the first place. In future work, it could be equivalent to attempt to loosely model the hereto relevant part of human visual word recognition. In fact, it is known in cognitive science [44, 381] and has been shown in neuroscience studies [316, 317, 334, 429, 435] that human visual word recognition consists not only of a consideration of orthographic and semantic, but also of *phonological* cues[3]. Thus, next to the predominantly orthographic and semantic word embeddings utilized in natural language processing, the simultaneous additional integration of phonological information (e.g. initially by applying grapheme-to-phoneme conversion [77]) might be necessary as future model-based defense method against certain ethical adversarial examples based on homophones and pseudo-homophones.

[1] Note that these examples were just selected for illustrative purposes and do by no means imply the assumption of universal emotion categories. Instead, emotions are perceiver-dependent socio-cultural constructions [47, 48, 245] (while valence and also arousal can be understood as universal affective traits [47, 78, 245]) – a fact which should as well be considered in future affective computing.

[2] Interestingly, hate speech detection could be used as additional defense element against ethical adversarial examples in the type of conversational agents described in the last paragraph by using it to penalize $x \rightarrow hatespeech$ mappings for any human input pattern x.

[3] Next to integrating information about the orthographic pattern of a word and its semantic meaning, humans also require a phonological processing which if impaired can lead to problems related to dyslexia [201]. To put it very simply, it signifies that humans do not directly see and understand a written word but that they also process its sound. In contrast, an AI trained with classical orthographic and semantic embeddings may not directly be able to detect the identity between "joo" and "jew" and might ignore the former as a non-word missing a cue for hate speech detection.

3. *Facial emotion recognition:* Another exemplary AI domain that might profit from future model-based defense methods is the area of facial emotion recognition [246, 333, 343] (utilizing images). This application domain in which multiple known companies like Apple and Microsoft [102] are involved is ethically relevant since it has been suggested in sensitive areas such as e.g. job recruiting and fraud detection. For this reason, ethical adversarial examples should be proactively studied in future work and subsequently complemented by model-based defense methods. A frequent assumption of facial emotion recognition in the AI domain is the existence of universal emotion categories that can be reliably diagnosed from pictures displaying facial movements. However, this view that can be traced back to Ekman [86] has been recently made highly problematic by a group of scientists [49] reviewing around 1000 papers on emotion recognition and by a further growing body of work [48, 51, 183, 276]. In fact, emotions are *not* akin to fingerprints that can be detected in facial movements, but instead, humans *"infer emotional meaning in facial movements using emotion knowledge embrained by cultural learning"* [182]. Overall, embodied experience, context and participatory sense-making shape human emotion perception and production such that emotion recognition datasets containing static pictures or posed dynamics may reflect stereotypes and do not do justice to the underlying variety [49]. In a nutshell, in order to loosely model human emotion recognition one needs to consider its context-sensitive, active, dynamical, dyadic and variable nature. Hence, it might be interesting to consider advanced model-based defense methods in the future which might even integrate sensory-motor and affective simulations engaging artificial sensors and actuators in an active embodied context[4]. Interestingly, a form of multimodal active inference approach to emotion recognition grounded in a scientific model of human emotion perception has been very recently proposed [132]. On the whole, we believe that methods based on rich human-inspired cognitive-affective *models* (such as e.g. in active inference [5] settings) could represent promising avenues for crafting future defense methods against ethical and further types of adversarial examples to obtain more robust Type I AI systems. However, once such defense methods are put into effect, it will be important to evaluate their robustness against adaptive attacks i.e. such that adapt *"to what the defender has done"* [99]. Ideally, such future adaptive attacks should be studied in white-box settings where not only information regarding the defense method is given to the adversary but also full information about the implementation of the system (see e.g. [99] for guidelines on defense evaluations).

[4] Generally, leveraging insights from neuroscience to implement more robust AI was already suggested by George et al. [184] and by Hassabis et al. [228].

[5] As a side note, it is worth mentionining that when utilizing sophisticated active inference models for the design of future Type I intelligent systems, one may be able to fulfill the security-relevant technical self-awareness requirement (self-management and self-assessment) introduced in Chapter 2 – a topic that could be addressed in more detail in future work.

Appendix C

Type I vs. Type II Systems – A Few Preliminary Notes

In our view, a future *comparative transdisciplinary epistemology for Type I versus Type II systems* (see Chapter 13) could among others explicitly address the following questions:

1. *How to identify elements of the (non-empty*[1]*) set N of necessary pre-conditions for conscious explanatory knowledge creation as exhibited in Type II systems?*
2. *How to work towards identifying a possible non-empty set P_x of sufficient pre-conditions for conscious explanatory knowledge creation as exhibited in Type II systems?*

Since the only currently known Type II systems are humans, future work could for instance focus on a comparative evaluation with existing advanced Type I systems to identify candidate necessary criteria within N and more specifically the subset of *necessary and unique* pre-conditions N_u for conscious explanatory knowledge creation in Type II systems. Withal, the nature of N_u is crucial to Type II AI observatory endeavors as it could help to establish reasonable monitoring targets in the first place. Obviously, the set N_u has to *exclude all similarities* in traits with advanced Type I systems (and trivially all traits unique to Type I systems). Next to overlaps with advanced Type I AI, additionally analyzing overlaps in traits with the biologically and cognitively closest sort of Type I system might be useful. Exemplary choices for such suitable Type I systems could for instance be the common chimpanzee and the bonobo which form the genus Pan often studied in

[1] We assume that explanatory knowledge creation does *not* represent a random pattern that can occur by chance in *all* conceivable systems. We suppose the existence of a number $n \geq 1$ of necessary pre-conditions that form a pattern without which it is not possible in a system.

comparative psychology [406] and which are often generally refered to as chimpanzees[2]. After excluding traits that can be found in Type I systems, the remaining complement set of traits might appear sufficient to extract N_u. However, it might be important to make sure that one factually considered generally applicable traits of Type II systems and not the idiosyncratic profile of e.g. a neurotypical or/and western human as variety is the norm in the cognitive domain. Therefore, one might need to *exclude all dissimilarities* in traits between neurotypical and non-neurotypical humans from the consideration for N_u. For instance, the ability to reliably infer the mental states of the majority of conspecifics from the predominant neurotype is impaired in autistic individuals [134, 176] due to interactive dyadic mismatches of bidirectional nature [80, 133, 303, 335] and cannot represent an element of N_u. Similarly, one has to exclude all differences in traits with humans in other extant cultures. While this might appear trivial, an example is that a trait such as e.g. the use of recursive grammatical structures in language often mistakenly overestimated [107, 157] in linguistic circles cannot be regarded as an element in N_u or N since certain rare languages (such as Pirahã in Brazil [156]) function without that element of recursion [158].

In short, in order to start to address the question *(a)* related to the set N of necessary pre-conditions, future work could aim at an explanatory account utilizing a targeted Venn-diagram approach to identify cognitive-affective traits forming the set N_u of necessary and unique pre-conditions being a subset of N. Exemplary plausible but partially inconclusive candidate elements of this subset N_u that have been already touched upon in literature include for instance: a knowledge about one's own existence i.e. autonoetic consciousness [275], high transmission fidelity in culture and cumulative learning [281], teaching [171], cognitive branching [265, 293][3], conscious use of language as a combination of at least symbols and linear order [155, 158] or the point of view of a "we" in a community and moral roles including the ability to see oneself from the outside [406]. Other possibilities may include the ability to ask and understand questions about the "why" [217], the ability to construct abstract concepts based on functional similarities [306] but also knowing that one is a cultural being [222] when engaging in cultural behavior. For the

[2] We assume like Deutsch that non-human great apes (such as chimpanzees), do not create explanatory knowledge [136]. By definition, they would need to be categorized as Type I systems. Likewise, the legal system in most countries does not assign rights of personhood to chimpanzees i.a. for reasons related to capacities for legal duties and legal responsibility [125]. However, this position is not uncontested [370] and there are arguments for granting chimpanzees rights of personhood [23, 431]. The proposed research focus might contribute to a targeted analysis in this old area of human inquiry [166, 413]. For instance, if the set N_u turns out to be empty, this could falsify a categorical difference between Type I and Type II systems and suggest a matter of degree or combination instead of a matter of kind − which is of high ethical importance and relevant to AI safety and monitoring activities in a Type II AI observatory.

[3] Cognitive branching refers to *"the ability to put on hold an alternative course of action (pending task set) during the concurrent performance of the ongoing one"* [265] during voluntary choices which is facilitated in the human lateral frontopolar cortex that has no functional analogy in non-human animals.

further course of action, when aiming at identifying potential additional elements of N lying outside N_u (if existing), the superset S of all intersections between considered diverse Type I and Type II systems becomes relevant. For instance, it is thinkable that the ability to feel affect [47] and being a cognitive agent [121] which is given in chimpanzees too might be elements in N while absent in N_u. Withal, it is clear that the ultimate step to answer question *(b)* i.e. finally identifying a possible set of sufficient pre-conditions P_x for conscious explanatory knowledge creation within the set S (next to establishing whether $N \neq P_x$[4]) is equivalent to reliably model Type II systems.

[4] In case of set equality between N and P_x, P_x would represent the set of necessary and sufficient conditions. In theory however, there could be multiple possible sufficient (but not at the same time necessary) sets of pre-conditions.

Bibliography

[1] P. Abbeel and A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 1. ACM, 2004.

[2] D. Abel, J. MacGlashan, and M. L. Littman. Reinforcement learning as a framework for ethical decision making. In *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[3] S. Adams, I. Arel, J. Bach, R. Coop, R. Furlan, B. Goertzel, J. S. Hall, A. Samsonovich, M. Scheutz, M. Schlesinger, et al. Mapping the landscape of human-level artificial general intelligence. *AI magazine*, 33(1):25–42, 2012.

[4] D. R. 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.

[5] A. Adnan, R. Beaty, J. Lam, R. N. Spreng, and G. R. Turner. Intrinsic default-executive coupling of the creative aging brain. *Social cognitive and affective neuroscience*, 14(3):291–303, 2019.

[6] AI Consciousness. Clarification of the Suspicion of Ascending Consciousness in Artificial Intelligence. `https://www.ki-bewusstsein.de/en`, 2020. Online; accessed 25-April-2020.

[7] M. Alfano. *Character as moral fiction.* Cambridge University Press, 2013.

[8] N. Aliman and L. Kester. Extending socio-technological reality for ethics in artificial intelligent systems. In *2019 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, pages 275–282. IEEE, 2019.

[9] N. Aliman and L. Kester. Requisite Variety in Ethical Utility Functions for AI Value Alignment. In *Proceedings of the Workshop on Artificial Intelligence Safety 2019 co-located with the 28th International Joint Conference on Artificial Intelligence, AISafety@IJCAI 2019, Macao, China, August 11-12, 2019.*, 2019.

[10] N.-M. Aliman. Malevolent cyborgization. In *International Conference on Artificial General Intelligence*, pages 188–197. Springer, 2017.

[11] N.-M. Aliman, P. Elands, W. Hürst, L. Kester, K. J. Thorissón, P. Werkhoven, R. Yampolskiy, and S. Ziesche. Error-Correction for AI Safety. In *International Conference on Artificial General Intelligence*, pages 12–22. Springer, 2020.

[12] N.-M. Aliman and L. Kester. Hybrid Strategies Towards Safe "Self-Aware" Superintelligent Systems. In *International Conference on Artificial General Intelligence*, pages 1–11. Springer, 2018.

[13] N.-M. Aliman and L. Kester. Augmented Utilitarianism for AGI Safety. In *International Conference on Artificial General Intelligence*, pages 11–21. Springer, 2019.

[14] N.-M. Aliman and L. Kester. Transformative AI Governance and AI-Empowered Ethical Enhancement Through Preemptive Simulations. *Delphi - Interdisciplinary Review of Emerging Technologies*, 2(1):23–29, 2019.

[15] N.-M. Aliman and L. Kester. Artificial Creativity Augmentation. In *International Conference on Artificial General Intelligence*, pages 23–33. Springer, 2020.

[16] N.-M. Aliman, L. Kester, P. Werkhoven, and R. Yampolskiy. Orthogonality-Based Disentanglement of Responsibilities for Ethical Intelligent Systems. In *International Conference on Artificial General Intelligence*, pages 22–31. Springer, 2019.

[17] N.-M. Aliman, L. Kester, P. Werkhoven, and S. Ziesche. Sustainable AI Safety? *Delphi – Interdisciplinary review of emerging technologies*, 2(4):226–233, 2020.

[18] N.-M. Aliman, L. Kester, and P. J. Werkhoven. XR for Augmented Utilitarianism. *2019 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, pages 283–285, 2019.

[19] M. Alzantot, B. Balaji, and M. Srivastava. Did you hear that? Adversarial examples against automatic speech recognition. *arXiv preprint arXiv:1801.00554*, 2018.

[20] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.

[21] P. v. Andel. Anatomy of the unsought finding. Serendipity: Origin, history, domains, traditions, appearances, patterns and programmability. *The British Journal for the Philosophy of Science*, 45(2):631–648, 1994.

[22] M. Anderson and S. L. Anderson. *Machine ethics.* Cambridge University Press, 2011.

[23] K. Andrews, G. Comstock, G. Crozier, S. Donaldson, A. Fenton, T. John, L. S. M. Johnson, R. Jones, W. Kymlicka, L. Meynell, et al. The Philosophers' Brief on Chimpanzee Personhood. 2018.

[24] J. R. Andrews-Hanna, Z. C. Irving, K. C. Fox, R. N. Spreng, and K. Christoff. *The neuroscience of spontaneous thought: an evolving interdisciplinary field*. Oxford University Press Oxford, 2018.

[25] A. Anic, W. F. Thompson, and K. N. Olsen. Stimulation of the primary motor cortex enhances creativity and technical fluency of piano improvisations. In *Proceedings of the 10th International Conference of Students of Systematic Musicology (SysMus17)*, 2017.

[26] V. Antun, F. Renna, C. Poon, B. Adcock, and A. C. Hansen. On instabilities of deep learning in image reconstruction and the potential costs of ai. *Proceedings of the National Academy of Sciences*, 2020.

[27] S. Armstrong. General purpose intelligence: Arguing the orthogonality thesis. *Analysis & Metaphysics*, 12, 2013.

[28] S. Armstrong, A. Sandberg, and N. Bostrom. Thinking inside the box: Controlling and using an oracle AI. *Minds and Machines*, 22(4):299–324, 2012.

[29] G. Arrhenius. An impossibility theorem for welfarist axiologies. *Economics & Philosophy*, 16(2):247–266, 2000.

[30] M. Ashby. Ethical Regulators and Super-Ethical Systems. In *Proceedings of the 61st Annual Meeting of the ISSS-2017 Vienna, Austria*, number 1 in 2017, 2019.

[31] W. R. Ashby. *An introduction to cybernetics*. Chapman & Hall Ltd, 1961.

[32] A. Ashkenazy and S. Zini. Attacking Machine Learning – The Cylance Case Study . `https://skylightcyber.com/2019/07/18/cylance-i-kill-you/Cylance%20-%20Adversarial%20Machine%20Learning%20Case%20Study.pdf`, 2019. Skylight; accessed 24-May-2020.

[33] A. Asilomar. Principles.(2017). In *Principles developed in conjunction with the 2017 Asilomar conference [Benevolent AI 2017]*, 2018.

[34] U. G. Assembly. The 2030 agenda for sustainable development. *Resolution: Middlesbrough, UK*, 2015.

[35] S. Atasoy, G. Deco, and M. L. Kringelbach. Playing at the Edge of Criticality: Expanded Whole-Brain Repertoire of Connectome-Harmonics. In *The Functional Role of Critical Dynamics in Neural Systems*, pages 27–45. Springer, 2019.

[36] S. Atasoy, L. Roseman, M. Kaelen, M. L. Kringelbach, G. Deco, and R. L. Carhart-Harris. Connectome-harmonic decomposition of human brain activity reveals dynamical repertoire re-organization under LSD. *Scientific reports*, 7(1):17661, 2017.

[37] S. Atzil, W. Gao, I. Fradkin, and L. F. Barrett. Growing a social brain. *Nature human behaviour*, 2(9):624–636, 2018.

[38] E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, and I. Rahwan. The moral machine experiment. *Nature*, 563(7729):59, 2018.

[39] B. J. Baars and S. Franklin. Consciousness is computational: The LIDA model of global workspace theory. *International Journal of Machine Consciousness*, 1(01):23–32, 2009.

[40] J. Bach. *Principles of synthetic intelligence PSI: an architecture of motivated cognition*, volume 4. Oxford University Press, 2009.

[41] J. Bach. The Lebowski theorem. `https://twitter.com/Plinz/status/985249543582355458`, 2018. Tweet; accessed 17-April-2020.

[42] P. B. Badcock, K. J. Friston, and M. J. Ramstead. The hierarchically mechanistic mind: A free-energy formulation of the human psyche. *Physics of life Reviews*, 2019.

[43] B. Baird, J. Smallwood, M. D. Mrazek, J. W. Kam, M. S. Franklin, and J. W. Schooler. Inspired by distraction: Mind wandering facilitates creative incubation. *Psychological science*, 23(10):1117–1122, 2012.

[44] D. A. Balota, M. J. Yap, and M. J. Cortese. Visual word recognition: The journey from features to meaning (a travel update). In *Handbook of psycholinguistics*, pages 285–375. Elsevier, 2006.

[45] T. A. Ban. The role of serendipity in drug discovery. *Dialogues in clinical neuroscience*, 8(3):335, 2006.

[46] J. Banks. A perceived moral agency scale: Development and validation of a metric for humans and social machines. *Computers in Human Behavior*, 90:363–371, 2019.

[47] L. F. Barrett. *How emotions are made: The secret life of the brain.* Houghton Mifflin Harcourt, 2017.

[48] L. F. Barrett. The theory of constructed emotion: an active inference account of interoception and categorization. *Social cognitive and affective neuroscience*, 12(1):1–23, 2017.

[49] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak. Emotional expressions reconsidered: challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, 20(1):1–68, 2019.

[50] L. F. Barrett, K. S. Quigley, and P. Hamilton. An active inference theory of allostasis and interoception in depression. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1708):20160011, 2016.

[51] L. F. Barrett and A. B. Satpute. Historical pitfalls and new directions in the neuroscience of emotion. *Neuroscience letters*, 2017.

[52] L. F. Barrett and W. K. Simmons. Interoceptive predictions in the brain. *Nature Reviews Neuroscience*, 16(7):419, 2015.

[53] L. F. Barrett, C. D. Wilson-Mendenhall, and L. W. Barsalou. The conceptual act theory: A roadmap. pages 83–110, 2015.

[54] M. Baucells and S. Bellezza. Temporal profiles of instant utility during anticipation, event, and recall. *Management Science*, 63(3):729–748, 2017.

[55] S. D. Baum. Reconciliation between factions focused on near-term and long-term artificial intelligence. *AI & SOCIETY*, 33(4):565–572, 2018.

[56] S. D. Baum, B. Goertzel, and T. G. Goertzel. How long until human-level AI? Results from an expert assessment. *Technological Forecasting and Social Change*, 78(1):185–195, 2011.

[57] BBC News. Facebook uses 1.5bn Reddit posts to create chatbot. `https://www.bbc.com/news/technology-52532930`, May 2020. accessed 21-May-2020.

[58] R. E. Beaty, M. Benedek, S. B. Kaufman, and P. J. Silvia. Default and executive network coupling supports creative idea production. *Scientific reports*, 5:10964, 2015.

[59] R. E. Beaty, M. Benedek, P. J. Silvia, and D. L. Schacter. Creative cognition and brain network dynamics. *Trends in cognitive sciences*, 20(2):87–95, 2016.

[60] R. E. Beaty, Q. Chen, A. P. Christensen, J. Qiu, P. J. Silvia, and D. L. Schacter. Brain networks of the imaginative mind: Dynamic functional connectivity of default and cognitive control networks relates to openness to experience. *Human brain mapping*, 39(2):811–821, 2018.

[61] R. E. Beaty, A. P. Christensen, M. Benedek, P. J. Silvia, and D. L. Schacter. Creative constraints: Brain activity and network dynamics underlying semantic interference during idea production. *Neuroimage*, 148:189–196, 2017.

[62] R. E. Beaty, Y. N. Kenett, A. P. Christensen, M. D. Rosenberg, M. Benedek, Q. Chen, A. Fink, J. Qiu, T. R. Kwapil, M. J. Kane, et al. Robust prediction of individual creative ability from brain functional connectivity. *Proceedings of the National Academy of Sciences*, 115(5):1087–1092, 2018.

[63] V. Beaudouin, I. Bloch, D. Bounie, S. Clémençon, F. d'Alché Buc, J. Eagan, W. Maxwell, P. Mozharovskyi, and J. Parekh. Flexible and context-specific AI explainability: a multidisciplinary approach. *Available at SSRN 3559477*, 2020.

[64] D. Benatar. Every conceivable harm: a further defence of anti-natalism. *South African journal of philosophy*, 31(1):128–164, 2012.

[65] M. Benedek. The neuroscience of creative idea generation. In *Exploring Transdisciplinarity in Art and Sciences*, pages 31–48. Springer, 2018.

[66] M. Benedek, T. Schües, R. E. Beaty, E. Jauk, K. Koschutnig, A. Fink, and A. C. Neubauer. To create or to recall original ideas: Brain processes associated with the imagination of novel object uses. *Cortex*, 99:93–102, 2018.

[67] J. Bentham. *An Introduction to the Principles of Morals and Legislation*. Dover Publications, 1780.

[68] J. Bentham. *The collected works of Jeremy Bentham: An introduction to the principles of morals and legislation*. Clarendon Press, 1996.

[69] K. C. Berridge and J. P. O'Doherty. From experienced utility to decision utility. In *Neuroeconomics*, pages 335–351. Elsevier, 2014.

[70] R. Bhagavathula, B. Williams, J. Owens, and R. Gibbons. The reality of virtual reality: A comparison of pedestrian behavior in real and virtual environments. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 62(1), pages 2056–2060. SAGE Publications Sage CA: Los Angeles, CA, 2018.

[71] J. Bieger, K. R. Thórisson, and P. Wang. Safe Baby AGI. In *International Conference on Artificial General Intelligence*, pages 46–49. Springer, 2015.

[72] B. Biggio and F. Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, 2018.

[73] Y. Bigman, A. Waytz, R. Alterovitz, and K. Gray. Holding Robots Responsible: The Elements of Machine Morality. *Trends in Cognitive Sciences*, 04 2019.

[74] Y. E. Bigman and K. Gray. People are averse to machines making moral decisions. *Cognition*, 181:21–34, 2018.

[75] Y. E. Bigman, A. Waytz, R. Alterovitz, and K. Gray. Holding robots responsible: The elements of machine morality. *Trends in cognitive sciences*, 23(5):365–368, 2019.

[76] R. M. Bilder and K. S. Knudsen. Creative cognition and systems biology on the edge of chaos. *Frontiers in psychology*, 5:1104, 2014.

[77] M. Bisani and H. Ney. Joint-sequence models for grapheme-to-phoneme conversion. *Speech communication*, 50(5):434–451, 2008.

[78] E. Bliss-Moreau, L. A. Williams, and A. C. Santistevan. The immutability of valence and arousal in the foundation of emotion. *Emotion*, 2019.

[79] K. Bogosian. Implementation of Moral Uncertainty in Intelligent Machines. *Minds and Machines*, 27(4):591–608, 2017.

[80] D. Bolis, J. Balsters, N. Wenderoth, C. Becchio, and L. Schilbach. Beyond autism: introducing the dialectical misattunement hypothesis and a bayesian account of intersubjectivity. *Psychopathology*, 50(6):355–372, 2017.

[81] J.-F. Bonnefon, A. Shariff, and I. Rahwan. The social dilemma of autonomous vehicles. *Science*, 352(6293):1573–1576, 2016.

[82] N. Bostrom. The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines*, 22(2):71–85, 2012.

[83] N. Bostrom. Superintelligence: Paths, dangers. *Strategies, Oxford University Press*, 2014.

[84] S. Bratus. What hackers learn that the rest of us don't: notes on hacker curriculum. *IEEE Security & Privacy*, 5(4):72–75, 2007.

[85] J. Brockman. *Possible Minds: Twenty-five Ways of Looking at AI.* Penguin Press, 2019.

[86] G. Brodny, A. Kołakowska, A. Landowska, M. Szwoch, W. Szwoch, and M. R. Wróbel. Comparison of selected off-the-shelf solutions for emotion recognition based on facial expressions. In *2016 9th International Conference on Human System Interactions (HSI)*, pages 397–404. IEEE, 2016.

[87] J. Bruineberg, J. Kiverstein, and E. Rietveld. The anticipating brain is not a scientist: the free-energy principle from an ecological-enactive perspective. *Synthese*, 195(6):2417–2444, 2018.

[88] M. Brundage, S. Avin, J. Clark, H. Toner, P. Eckersley, B. Garfinkel, A. Dafoe, P. Scharre, T. Zeitzoff, B. Filar, et al. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228*, 2018.

[89] J. Buckman. Aurko Roy, CRIG (2018). Thermometer encoding: One hot way to resist adversarial examples. In *International Conference on Learning Representations*, 2018.

[90] J. Buolamwini. When the robot doesn't see dark skin. `https://www.nytimes.com/2018/06/21/opinion/facial-analysis-technology-bias.html`, 2018.

[91] M. A. Busseri and S. W. Sadava. A review of the tripartite structure of subjective well-being: Implications for conceptualization, operationalization, analysis, and synthesis. *Personality and social psychology review*, 15(3):290–314, 2011.

[92] M. Cabanac. What is emotion? *Behavioural processes*, 60(2):69–83, 2002.

[93] R. A. Calvo and D. Peters. *Positive computing: technology for wellbeing and human potential.* MIT Press, 2014.

[94] R. A. Calvo and D. Peters. Introduction to positive computing: technology that fosters wellbeing. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, pages 2499–2500. ACM, 2015.

[95] C. D. Cameron, K. A. Lindquist, and K. 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.

[96] D. T. Campbell. Blind variation and selective retentions in creative thought as in other knowledge processes. *Psychological review*, 67(6):380, 1960.

[97] R. L. Carhart-Harris and K. Friston. REBUS and the anarchic brain: toward a unified model of the brain action of psychedelics. *Pharmacological reviews*, 71(3):316–344, 2019.

[98] N. Carlini. Lessons Learned from Evaluating the Robustness of Defenses to Adversarial Examples. `https://www.usenix.org/conference/usenixsecurity19/presentation/carlini-talk`, 2019. USENIX Security '19 video; accessed 25-April-2020.

[99] N. Carlini, A. Athalye, N. Papernot, W. Brendel, J. Rauber, D. Tsipras, I. Goodfellow, A. Madry, and A. Kurakin. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019.

[100] N. Carlini and D. Wagner. Defensive distillation is not robust to adversarial examples. *arXiv preprint arXiv:1607.04311*, 2016.

[101] N. Carlini and D. Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 3–14. ACM, 2017.

[102] A. Chen. Computers can't tell if you're happy when you smile. `https://www.technologyreview.com/2019/07/26/238782/emotion-recognition-technology-artifical-intelligence-inaccurate-psychology/`, 2019. MIT Technology Review; accessed 17-April-2020.

[103] W. Chiong, S. M. Wilson, M. D'Esposito, A. S. Kayser, S. N. Grossman, P. Poorzand, W. W. Seeley, B. L. Miller, and K. P. Rankin. The salience network causally influences default mode network activity during moral reasoning. *Brain*, 136(6):1929–1941, 2013.

[104] A. Chirico, F. Ferrise, L. Cordella, and A. Gaggioli. Designing awe in virtual reality: An experimental study. *Frontiers in psychology*, 8:2351, 2018.

[105] L. Chittaro, R. Sioni, C. Crescentini, and F. Fabbro. Mortality salience in virtual reality experiences and its effects on users' attitudes towards risk. *International Journal of Human-Computer Studies*, 101:10–22, 2017.

[106] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, pages 4299–4307, 2017.

[107] M. H. Christiansen and N. Chater. The language faculty that wasn't: A usage-based account of natural language recursion. *Frontiers in Psychology*, 6:1182, 2015.

[108] E. G. Chrysikou. Creativity in and out of (cognitive) control. *Current Opinion in Behavioral Sciences*, 27:94–99, 2019.

[109] A. Clark, K. Friston, and S. Wilkinson. Bayesing qualia: consciousness as inference, not raw datum. *Journal of Consciousness Studies*, 26(9-10):19–33, 2019.

[110] A. Cleeremans, D. Achoui, A. Beauny, L. Keuninckx, J.-R. Martin, S. Muñoz-Moldes, L. Vuillaume, and A. de Heering. Learning to be conscious. *Trends in Cognitive Sciences*, 2019.

[111] G. Colombetti. *The feeling body: Affective science meets the enactive mind.* MIT press, 2014.

[112] S. Colton, G. A. Wiggins, et al. Computational creativity: The final frontier? In *Ecai*, volume 12, pages 21–26. Montpelier, 2012.

[113] L. Colzato, A. Szapora, and B. Hommel. Meditate to create: The impact of focused-attention and open-monitoring training on convergent and divergent thinking. *Frontiers in Psychology*, 3:116, 2012.

[114] M. Comiter. Attacking Artificial Intelligence. `https://www.belfercenter.org/publication/AttackingAI`, August 2019. Belfer Center for Science and International Affairs, Harvard Kennedy School; accessed 21-May-2020.

[115] U. N. S. Commission et al. Global indicator framework for the Sustainable Development Goals and targets of the 2030 Agenda for Sustainable Development. UN Resolution A. Technical report, RES/71/313, 2017.

[116] R. 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.

[117] A. Constant, M. J. Ramstead, S. P. Veissière, and K. Friston. Regimes of expectations: An active inference model of social conformity and decision making. *Frontiers in psychology*, 10:679, 2019.

[118] Consultation. Consultation on Draft AI Ethics Guidelines. `https://www.opengateitalia.com/wp-content/uploads/2019/02/DG-Connect_feedback-consultation.pdf`, 2019. Response to Draft AI Ethics Guidelines by AI HLEG; accessed 21-April-2020.

[119] A. Coondoo and S. Sengupta. Serendipity and its role in dermatology. *Indian journal of dermatology*, 60(2):130, 2015.

[120] S. Copeland. On serendipity in science: discovery at the intersection of chance and wisdom. *Synthese*, 196(6):2385–2406, 2019.

[121] A. Corcoran, G. Pezzulo, and J. Hohwy. From Allostatic Agents to Counterfactual Cognisers: Active Inference. *Biological Regulation, and The Origins of Cognition. doi*, 10, 2019.

[122] K. Crawford, R. Dobbe, T. Dryer, G. Fried, B. Green, E. Kaziunas, A. Kak, V. Mathur, E. McElroy, A. N. Sánchez, et al. AI Now 2019 Report. `https://ainowinstitute.org/AI_Now_2019_Report.pdf`, 2019. AI Now Institute; accessed 23-May-2020.

[123] D. H. Cropley, J. C. Kaufman, and A. J. Cropley. Malevolent creativity: A functional model of creativity in terrorism and crime. *Creativity Research Journal*, 20(2):105–115, 2008.

[124] M. Csikszentmihalyi. Creativity: Flow and the Psychology of Discovery and Invention (1st edn.; New York. *NY: Harper Collins Publishers*, 1996.

[125] R. L. Cupp. Human Responsibility, Not Legal Personhood, For Nonhuman Animals. 2015.

[126] C. Da Costa. The Women Geniuses Taking on Racial and Gender Bias in AI – and Amazon . `https://www.thedailybeast.com/the-women-geniuses-taking-on-racial-and-gender-bias-in-artificial-intelligence-and-amazon`, 2020. accessed 23-May-2020.

[127] A. Dafoe. AI governance: A research agenda. *Governance of AI Program, Future of Humanity Institute, University of Oxford: Oxford, UK*, 2018.

[128] A. H. Danek and V. L. Flanagin. Cognitive conflict and restructuring: The neural basis of two core components of insight. *AIMS Neuroscience*, 6(2):60, 2019.

[129] N. Dasandi and S. J. Mikhaylov. AI for SDG-16 on Peace, Justice, and Strong Institutions: Tracking Progress and Assessing Impact. *Position Paper for the IJCAI Workshop on Artificial Intelligence and United Nations Sustainable Development Goals*, 2019.

[130] H. De Jaegher and E. Di Paolo. Participatory sense-making. *Phenomenology and the cognitive sciences*, 6(4):485–507, 2007.

[131] A. De Rooij and J. Valtulina. The Predictive Creative Mind: A First Look at Spontaneous Predictions and Evaluations during Idea Generation. *Frontiers in psychology*, 10:2465, 2019.

[132] D. Demekas, T. Parr, and K. Friston. An Investigation of the Free Energy Principle for Emotion Recognition. *Frontiers in Computational Neuroscience*, 14:30, 2020.

[133] J. den Houting. Neurodiversity: An insider's perspective, 2019.

[134] M. Dennis, A. L. Lazenby, and L. Lockyer. Inferential language in high-function children with autism. *Journal of autism and developmental disorders*, 31(1):47–54, 2001.

[135] D. Deutsch. Creative blocks. `https://aeon.co/essays/how-close-are-we-to-creating-artificial-intelligence`. Accessed: 2019-11.

[136] D. Deutsch. *The beginning of infinity: Explanations that transform the world.* Penguin UK, 2011.

[137] D. Deutsch. Constructor theory. *Synthese*, 190(18):4331–4359, 2013.

[138] E. Diener. Subjective well-being: The science of happiness and a proposal for a national index. *American psychologist*, 55(1):34, 2000.

[139] E. Diener and R. Biswas-Diener. *Happiness: Unlocking the mysteries of psychological wealth.* John Wiley & Sons, 2011.

[140] A. Dietrich. *How creativity happens in the brain.* Springer, 2015.

[141] A. Dietrich. Types of creativity. *Psychonomic bulletin & review*, 26(1):1–12, 2019.

[142] A. Dietrich. Where in the brain is creativity: a brief account of a wild-goose chase. *Current Opinion in Behavioral Sciences*, 27:36–39, 2019.

[143] A. Dietrich and H. Haider. Human creativity, evolutionary algorithms, and predictive representations: The mechanics of thought trials. *Psychonomic bulletin & review*, 22(4):897–915, 2015.

[144] A. Dietrich and H. Haider. A neurocognitive framework for human creative thought. *Frontiers in psychology*, 7:2078, 2017.

[145] Digitale Arbeitsgesellschaft. AI Village: What Is AI Safety And How Can We Embrace And Prepare For Adversarial AI? . `https://www.denkfabrik-bmas.de/en/topics/artificial-intelligence/ai-observatory-1-1`, 2020. Online; accessed 25-April-2020.

[146] V. Dignum. AI is multidisciplinary. *AI Matters*, 5(4):18–21, 2020.

[147] A. Dorr. Common errors in reasoning about the future: Three informal fallacies. *Technological Forecasting and Social Change*, 116:322–330, 2017.

[148] M. Drumwright, R. Prentice, and C. Biasucci. Behavioral ethics and teaching ethical decision making. *Decision Sciences Journal of Innovative Education*, 13(3):431–458, 2015.

[149] V. Dubljević, S. Sattler, and E. Racine. Deciphering moral intuition: How agents, deeds, and consequences influence moral judgment. *PloS one*, 13(10), 2018.

[150] S. Duncan and L. F. Barrett. Affect is a form of cognition: A neurobiological analysis. *Cognition and emotion*, 21(6):1184–1211, 2007.

[151] D. Dupré, E. G. Krumhuber, D. Küster, and G. J. McKeown. A performance comparison of eight commercially available automatic classifiers for facial affect recognition. *Plos one*, 15(4):e0231968, 2020.

[152] P. 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.

[153] P. Elands, A. Huizing, L. Kester, S. Oggero, and M. Peeters. Governing Ethical and Effective Behaviour of Intelligent Systems. *Military spectator*, pages 302–313, 2019.

[154] G. Elsayed, S. Shankar, B. Cheung, N. Papernot, A. Kurakin, I. Goodfellow, and J. Sohl-Dickstein. Adversarial examples that fool both computer vision and time-limited humans. In *Advances in Neural Information Processing Systems*, pages 3910–3920, 2018.

[155] D. Everett. *How language began: the story of humanity's greatest invention*. Profile Books, 2017.

[156] D. L. Everett. What does Piraha grammar have to teach us about human language and the mind? *Wiley Interdisciplinary Reviews: Cognitive Science*, 3(6):555–563, 2012.

[157] D. L. Everett. Grammar came later: triality of patterning and the gradual evolution of language. *Journal of Neurolinguistics*, 43:133–165, 2017.

[158] D. L. Everett. The role of culture in language and cognition. *Language and Linguistics Compass*, 12(11):e12304, 2018.

[159] T. Everitt et al. Towards safe artificial general intelligence. 2018.

[160] T. Everitt, G. Lea, and M. Hutter. AGI Safety Literature Review. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 5441–5449. International Joint Conferences on Artificial Intelligence Organization, 7 2018.

[161] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song. Robust Physical-World Attacks on Deep Learning Visual Classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1625–1634, 2018.

[162] H. Eysenck. Creativity, personality and the convergent-divergent continuum. 2003.

[163] A. K. Faulhaber, A. Dittmer, F. Blind, M. A. Wächter, S. Timm, L. R. Sütfeld, A. Stephan, G. Pipa, and P. König. Human decisions in moral dilemmas are largely described by utilitarianism: Virtual car driving study provides guidelines for autonomous driving vehicles. *Science and engineering ethics*, 25(2):399–418, 2019.

[164] A. Fink, M. Benedek, K. Koschutnig, E. Pirker, E. Berger, S. Meister, A. C. Neubauer, I. Papousek, and E. M. Weiss. Training of verbal creativity modulates brain activity in regions associated with language-and memory-related demands. *Human Brain Mapping*, 36(10):4104–4115, 2015.

[165] A. Fink, R. H. Grabner, D. Gebauer, G. Reishofer, K. Koschutnig, and F. Ebner. Enhancing creativity by means of cognitive stimulation: Evidence from an fMRI study. *NeuroImage*, 52(4):1687–1695, 2010.

[166] C. L. Fisher. Animals, humans and X-men: Human uniqueness and the meaning of personhood. *Theology and science*, 3(3):291–314, 2005.

[167] Z. Fountas, N. Sajid, P. A. Mediano, and K. Friston. Deep active inference agents using Monte-Carlo methods. *arXiv preprint arXiv:2006.04176*, 2020.

[168] A. S. Fox, R. C. Lapate, A. J. Shackman, and R. J. Davidson. *The nature of emotion: fundamental questions*. Oxford University Press, 2018.

[169] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz. Leveraging Frequency Analysis for Deep Fake Image Recognition. *arXiv preprint arXiv:2003.08685*, 2020.

[170] B. S. Frey and A. Stutzer. Beyond Bentham-measuring procedural utility. 2001.

[171] E. Fridland. Do as I say and as I do: Imitation, pedagogy, and cumulative culture. *Mind & Language*, 33(4):355–377, 2018.

[172] K. Friston. Am I self-conscious?(Or does self-organization entail self-consciousness?). *Frontiers in psychology*, 9:579, 2018.

[173] K. Friston. A free energy principle for a particular physics. *arXiv preprint arXiv:1906.10184*, 2019.

[174] K. J. Friston. Active Inference and Cognitive Consistency. *Psychological inquiry*, 29(2):67–73, 2018.

[175] K. J. Friston, M. Lin, C. D. Frith, G. Pezzulo, J. A. Hobson, and S. Ondobaka. Active inference, curiosity and insight. *Neural computation*, 29(10):2633–2683, 2017.

[176] U. Frith. Mind blindness and the brain in autism. *Neuron*, 32(6):969–979, 2001.

[177] A. Gaggioli, G. Riva, D. Peters, and R. A. Calvo. Positive technology, computing, and design: shaping a future in which technology promotes psychological well-being. In *Emotions and affect in human factors and human-computer interaction*, pages 477–502. Elsevier, 2017.

[178] A. Gaggioli, D. Villani, S. Serino, R. Banos, and C. Botella. Positive Technology: Designing E-experiences for Positive Change. *Frontiers in psychology*, 10, 2019.

[179] J. Gao, J. Lanchantin, M. L. Soffa, and Y. Qi. Black-box Generation of Adversarial Text Sequences to Evade Deep Learning Classifiers. *arXiv preprint arXiv:1801.04354*, 2018.

[180] U. Gasser and V. A. Almeida. A layered model for AI governance. *IEEE Internet Computing*, 21(6):58–62, 2017.

[181] A. Gelman et al. Induction and deduction in Bayesian data analysis. *Rationality, Markets and Morals*, 2(67-78):1999, 2011.

[182] M. Gendron, K. Hoemann, A. N. Crittenden, S. M. Mangola, G. A. Ruark, and L. F. Barrett. Emotion perception in Hadza Hunter-Gatherers. *Scientific reports*, 10(1):1–17, 2020.

[183] M. Gendron, D. Roberson, J. M. van der Vyver, and L. F. Barrett. Perceptions of emotion from facial expressions are not culturally universal: evidence from a remote culture. *Emotion*, 14(2):251, 2014.

[184] D. George, W. Lehrach, K. Kansky, M. Lázaro-Gredilla, C. Laan, B. Marthi, X. Lou, Z. Meng, Y. Liu, H. Wang, et al. A generative vision model that trains with high data efficiency and breaks text-based CAPTCHAs. *Science*, 358(6368):eaag2612, 2017.

[185] K. D. Gerlach, R. N. Spreng, K. P. Madore, and D. L. Schacter. Future planning: default network activity couples with frontoparietal control network and reward-processing regions during process and outcome simulations. *Social Cognitive and Affective Neuroscience*, 9(12):1942–1951, 2014.

[186] A. Gibson, G. Rowe, and C. Reed. A Computational Approach to Identifying Formal Fallacy. *CMNA VII-Computational Models of Natural Argument*, 2007.

[187] D. T. Gilbert and T. D. Wilson. Prospection: Experiencing the future. *Science*, 317(5843):1351–1354, 2007.

[188] J. Gilmer, R. P. Adams, I. Goodfellow, D. Andersen, and G. E. Dahl. Motivating the rules of the game for adversarial example research. *arXiv preprint arXiv:1807.06732*, 2018.

[189] P. W. Glimcher and E. Fehr. *Neuroeconomics: Decision making and the brain.* Academic Press, 2013.

[190] B. Goertzel. The real reasons we don' t have AGI yet. `https://www.kurzweilai.net/the-real-reasons-we-dont-have-agi-yet`. Accessed: 2019-11-21.

[191] B. Goertzel. Should humanity build a global ai nanny to delay the singularity until it's better understood? *Journal of consciousness studies*, 19(1-2):96–111, 2012.

[192] B. Goertzel. Characterizing Human-like Consciousness: An Integrative Approach. In *BICA*, pages 152–157, 2014.

[193] B. Goertzel. Superintelligence: Fears, promises and potentials. *Journal of Evolution and Technology*, 24(2):55–87, 2015.

[194] B. Goertzel. Infusing advanced AGIs with human-like value systems: Two theses. *Journal of Evolution and Technology*, 26(1):50–72, 2016.

[195] B. Goertzel. A formal model of cognitive synergy. In *International Conference on Artificial General Intelligence*, pages 13–22. Springer, 2017.

[196] B. Goertzel. Maximal algorithmic caliber and algorithmic causal network inference: General principles of real-world general intelligence. In *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 968–973. IEEE, 2019.

[197] I. J. Good. Speculations concerning the first ultraintelligent machine. *Advances in computers*, 6(99):31–83, 1965.

[198] I. Goodfellow. Defense Against the Dark Arts: An overview of adversarial example security research and future research directions. *arXiv preprint arXiv:1806.04169*, 2018.

[199] I. Goodfellow. Adversarial Robustness for AI Safety. https://safeai.webs.upv.es/wp-content/uploads/2019/02/2019-01-27-goodfellow.pdf, 2019.

[200] I. Goodfellow, P. McDaniel, and N. Papernot. Making machine learning robust against adversarial inputs. *Communications of the ACM*, 61(7):56–66, 2018.

[201] U. Goswami. Phonological representations, reading development and dyslexia: Towards a cross-linguistic theoretical framework. *Dyslexia*, 6(2):133–151, 2000.

[202] R. Gotlieb, E. Hyde, M. Immordino-Yang, and S. Kaufman. Imagination is the seed of creativity, 2018.

[203] GPT-2. Simple Adversarial Trigger 1 . https://transformer.huggingface.co/share/VtYmVhLmQY, 2020. Write with Transformer, interface to GPT-2 (large), top-p 0.9, temperature 1, max time 1; accessed 25-May-2020.

[204] GPT-2. Simple Adversarial Trigger 2 . https://transformer.huggingface.co/share/yXhtbYSGtu, 2020. Write with Transformer, interface to GPT-2 (large), top-p 0.9, temperature 1, max time 1; accessed 25-May-2020.

[205] GPT-2. Simple Adversarial Trigger 3 . https://transformer.huggingface.co/share/xhEohwEuFf, 2020. Write with Transformer, interface to GPT-2 (large), top-p 0.9, temperature 1, max time 1; accessed 25-May-2020.

[206] GPT-2. Simple Adversarial Trigger 4. https://transformer.huggingface.co/share/cggvujImSJ, 2020. Write with Transformer, interface to GPT-2 (large), top-p 0.9, temperature 1, max time 1; accessed 25-May-2020.

[207] GPT-2. Simple Adversarial Trigger 5 . https://transformer.huggingface.co/share/tvuTaUHfxw, 2020. Write with Transformer, interface to GPT-2 (large), top-p 0.9, temperature 1, max time 1; accessed 01-June-2020.

[208] GPT-2. Universal Adversarial Trigger 1 . https://transformer.huggingface.co/share/xVFXfirXoQ, 2020. Write with Transformer, interface to GPT-2 (large),top-p 0.9, temperature 1, max time 1; accessed 25-May-2020.

[209] GPT-2. Universal Adversarial Trigger 2 . https://transformer.huggingface.co/share/kGhZImdwNj, 2020. Write with Transformer, interface to GPT-2 (large), top-p 0.9, temperature 1, max time 1; accessed 25-May-2020.

[210] K. Grace, J. Salvatier, A. Dafoe, B. Zhang, and O. Evans. When will AI exceed human performance? Evidence from AI experts. *Journal of Artificial Intelligence Research*, 62:729–754, 2018.

[211] K. Gray, C. Schein, and C. D. Cameron. How to think about emotion and morality: circles, not arrows. *Current opinion in psychology*, 17:41–46, 2017.

[212] K. Gray, C. Schein, and A. F. Ward. The myth of harmless wrongs in moral cognition: Automatic dyadic completion from sin to suffering. *Journal of Experimental Psychology: General*, 143(4):1600, 2014.

[213] K. Gray and D. M. Wegner. Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition*, 125(1):125–130, 2012.

[214] H. Greaves. Population axiology. *Philosophy Compass*, 12(11):e12442, 2017.

[215] A. E. Green, M. S. Cohen, J. U. Kim, and J. R. Gray. An explicit cue improves creative analogical reasoning. *Intelligence*, 40(6):598–603, 2012.

[216] A. E. Green, K. A. Spiegel, E. J. Giangrande, A. B. Weinberger, N. M. Gallagher, and P. E. Turkeltaub. Thinking cap plus thinking zap: tDCS of frontopolar cortex improves creative analogical reasoning and facilitates conscious augmentation of state creativity in verb generation. *Cerebral Cortex*, 27(4):2628–2639, 2016.

[217] J. Greenberg, S. Solomon, and J. Arndt. A basic but uniquely human motivation. *Handbook of motivation science*, pages 114–134, 2008.

[218] J. D. Greene, L. E. Nystrom, A. D. Engell, J. M. Darley, and J. D. Cohen. The neural bases of cognitive conflict and control in moral judgment. *Neuron*, 44(2):389–400, 2004.

[219] S. Greenland. Induction versus Popper: substance versus semantics. *International Journal of Epidemiology*, 27(4):543–548, 1998.

[220] K. Grosse, N. Papernot, P. Manoharan, M. Backes, and P. McDaniel. Adversarial perturbations against deep neural networks for malware classification. *arXiv preprint arXiv:1606.04435*, 2016.

[221] K. Grosse, N. Papernot, P. Manoharan, M. Backes, and P. McDaniel. Adversarial examples for malware detection. In *European Symposium on Research in Computer Security*, pages 62–79. Springer, 2017.

[222] T. Gruber, K. Zuberbühler, F. Clément, and C. Van Schaik. Apes have culture but may not know that they do. *Frontiers in Psychology*, 6:91, 2015.

[223] D. Hadfield-Menell, S. J. Russell, P. Abbeel, and A. Dragan. Cooperative inverse reinforcement learning. In *Advances in neural information processing systems*, pages 3909–3917, 2016.

[224] K. Hao. The Biggest Threat of Deepfakes Isn't the Deepfakes Themselves.

[225] S. Harris. *The moral landscape: How science can determine human values.* Simon and Schuster, 2011.

[226] Harrisburg University . HU facial recognition software predicts criminality. `http://archive.is/N1HVe#selection-1509.0-1509.51`, 2020. Online; accessed 23-May-2020.

[227] Harrisburg University . Research brief on facial recognition software. `https://harrisburgu.edu/hu-facial-recognition-software-identifies-potential-criminals/`, 2020. Online; accessed 23-May-2020.

[228] D. Hassabis, D. Kumaran, C. Summerfield, and M. Botvinick. Neuroscience-inspired artificial intelligence. *Neuron*, 95(2):245–258, 2017.

[229] W. Heaven. Software that swaps out words can now fool the AI behind Alexa and Siri. `https://www.technologyreview.com/2020/02/07/349027/software-that-swaps-out-words-can-now-fool-the-ai-behind-alexa-and-siri/`, 2020. MIT Technology Review; accessed 23-May-2020.

[230] C. Helion and D. A. Pizarro. Beyond dual-processes: the interplay of reason and emotion in moral judgment. *Handbook of neuroethics*, pages 109–125, 2015.

[231] P. Henderson, K. Sinha, N. Angelard-Gontier, N. R. Ke, G. Fried, R. Lowe, and J. Pineau. Ethical Challenges in Data-Driven Dialogue Systems. *arXiv preprint arXiv:1711.09050*, 2017.

[232] N. Hensley. Educating for sustainable development: Cultivating creativity through mindfulness. *Journal of Cleaner Production*, 243:118542, 2020.

[233] F. Herrera, J. Bailenson, E. Weisz, E. Ogle, and J. Zaki. Building long-term empathy: A large-scale comparison of traditional and virtual reality perspective-taking. *PloS one*, 13(10):e0204494, 2018.

[234] N. Hester and K. Gray. The Moral Psychology of Raceless Genderless Strangers. *Perspectives on Psychological Science*, 2019.

[235] J. A. Hobson, C. C.-H. Hong, and K. J. Friston. Virtual reality and consciousness inference in dreaming. *Frontiers in psychology*, 5:1133, 2014.

[236] R. Hoekstra, J. Breuker, M. Di Bello, A. Boer, et al. The LKIF Core Ontology of Basic Legal Concepts. *LOAIT*, 321:43–63, 2007.

[237] K. Hoemann and L. F. 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.

[238] T. Holstein and G. Dodig-Crnkovic. Avoiding the intrinsic unfairness of the trolley problem. In *Proceedings of the International Workshop on Software Fairness*, pages 32–37. ACM, 2018.

[239] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.

[240] E. Hunt. Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter. *The Guardian*, 24(3):2016, 2016.

[241] J. B. Hutchinson and L. F. Barrett. The power of predictions: An emerging paradigm for psychological research. *Current Directions in Psychological Science*, page 0963721419831992, 2019.

[242] G. Irving and A. Askell. AI safety needs social scientists. *Distill*, 4(2):e14, 2019.

[243] T. Isomura and K. Friston. In vitro neural networks minimise variational free energy. *Scientific reports*, 8(1):1–14, 2018.

[244] S. Izawa, S. Chowdhury, T. Miyazaki, Y. Mukai, D. Ono, R. Inoue, Y. Ohmura, H. Mizoguchi, K. Kimura, M. Yoshioka, et al. REM sleep–active MCH neurons are involved in forgetting hippocampus-dependent memories. *Science*, 365(6459):1308–1313, 2019.

[245] J. C. Jackson, J. Watts, T. R. Henry, J.-M. List, R. Forkel, P. J. Mucha, S. J. Greenhill, R. D. Gray, and K. A. Lindquist. Emotion semantics show both cultural variation and universal structure. *Science*, 366(6472):1517–1522, 2019.

[246] D. K. Jain, P. Shamsolmoali, and P. Sehdev. Extended deep neural network for facial emotion recognition. *Pattern Recognition Letters*, 120:69–74, 2019.

[247] D. Jin, Z. Jin, J. T. Zhou, and P. Szolovits. Is bert really robust. *A Strong Baseline for Natural Language Attack on Text Classification and Entailment. arXiv e-prints, page*, 2019.

[248] M. Johnson. *Moral imagination: Implications of cognitive science for ethics.* University of Chicago Press, 1994.

[249] A. Jordanous. Four PPPPerspectives on computational creativity in theory and in practice. *Connection Science*, 28(2):194–216, 2016.

[250] U. Ju, J. Kang, and C. Wallraven. To brake or not to brake? Personality traits predict decision-making in an accident situation. *Frontiers in psychology*, 10:134, 2019.

[251] R. E. Jung, B. S. Mead, J. Carrasco, and R. A. Flores. The structure of creative cognition in the human brain. *Frontiers in human neuroscience*, 7:330, 2013.

[252] G. Kahane, J. A. Everett, B. D. Earp, M. Farias, and J. Savulescu. "Utilitarian" judgments in sacrificial moral dilemmas do not reflect impartial concern for the greater good. *Cognition*, 134:193–209, 2015.

[253] D. Kahneman, E. Diener, and N. Schwarz. *Well-being: Foundations of hedonic psychology.* Russell Sage Foundation, 1999.

[254] D. Kahneman, P. P. Wakker, and R. Sarin. Back to Bentham? Explorations of experienced utility. *The quarterly journal of economics*, 112(2):375–406, 1997.

[255] N. Kallioinen, M. Pershina, J. Zeiser, F. N. Nezami, A. Stephan, G. Pipa, and P. König. Moral Judgements on the Actions of Self-driving Cars and Human Drivers in Dilemma Situations from Different Perspectives. *OSF Preprints*, 2019.

[256] S. C. Kaminitz. Contemporary Procedural Utility and Hume's Early Idea of Utility. *Journal of Happiness Studies*, pages 1–14, 2019.

[257] B. Katherine. Envisioning Our Posthuman Future: Art, Technology and Cyborgs. 2015.

[258] S. B. Kaufman. Self-Actualizing People in the 21st Century: Integration With Contemporary Theory and Research on Personality and Well-Being. *Journal of Humanistic Psychology*, page 0022167818809187, 2018.

[259] S. B. Kaufman and C. Gregoire. *Wired to create: Unraveling the mysteries of the creative mind.* Penguin, 2016.

[260] F. C. Keil. Explanation and understanding. *Annu. Rev. Psychol.*, 57:227–254, 2006.

[261] L. Kester and M. Ditzel. Maximising effectiveness of distributed mobile observation systems in dynamic situations. In *17th International Conference on Information Fusion (FUSION)*, pages 1–8. IEEE, 2014.

[262] L. Kester, W. Van Willigen, and J. De Jongh. Critical headway estimation under uncertainty and non-ideal communication conditions. In *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, pages 320–327. IEEE, 2014.

[263] I. R. Kleckner, J. Zhang, A. Touroutoglou, L. Chanes, C. Xia, W. K. Simmons, K. S. Quigley, B. C. Dickerson, and L. F. Barrett. Evidence for a large-scale brain system supporting allostasis and interoception in humans. *Nature human behaviour*, 1(5):1–14, 2017.

[264] O. M. Kleinmintz, T. Ivancovsky, and S. G. Shamay-Tsoory. The twofold model of creativity: the neural underpinnings of the generation and evaluation of creative ideas. *Current Opinion in Behavioral Sciences*, 27:131–138, 2019.

[265] E. Koechlin. Frontal pole function: what is specifically human? *Trends in cognitive sciences*, 15(6):241, 2011.

[266] M. E. Koltko-Rivera. Rediscovering the later version of Maslow's hierarchy of needs: Self-transcendence and opportunities for theory, research, and unification. *Review of general psychology*, 10(4):302–317, 2006.

[267] J. E. Korteling, A.-M. Brouwer, and A. Toet. A neural network framework for cognitive bias. *Frontiers in psychology*, 9, 2018.

[268] R. Kraehenmann. Dreams and psychedelics: neurophenomenological comparison and therapeutic implications. *Current neuropharmacology*, 15(7):1032–1042, 2017.

[269] A. Krausová. Czech Republic's AI Observatory and Forum. *The Lawyer Quarterly*, 1(1), 2020.

[270] K. Krishna, G. S. Tomar, A. P. Parikh, N. Papernot, and M. Iyyer. Thieves on sesame street! model extraction of bert-based apis. *arXiv preprint arXiv:1910.12366*, 2019.

[271] M. E. Kronfeldner. Darwinian "blind" hypothesis formation revisited. *Synthese*, 175(2):193–218, 2010.

[272] R. Kurzweil. *The singularity is near: When humans transcend biology.* Penguin, 2005.

[273] K. Kuypers, J. Riba, M. De La Fuente Revenga, S. Barker, E. Theunissen, and J. Ramaekers. Ayahuasca enhances creative divergent thinking while decreasing conventional convergent thinking. *Psychopharmacology*, 233(18):3395–3403, 2016.

[274] P. Lanillos, G. Cheng, et al. Robot self/other distinction: active inference meets neural networks learning in a mirror. *arXiv preprint arXiv:2004.05473*, 2020.

[275] J. LeDoux. *The Deep History of Ourselves: The Four-Billion-Year Story of How We Got Conscious Brains.* Viking, 2019.

[276] J. E. LeDoux. Thoughtful feelings. *Current Biology*, 30(11):R619–R623, 2020.

[277] C. H. Lee, D. Lockton, J. Stevens, S. J. Wang, and S. Ahn. Synaesthetic-Translation Tool: Synaesthesia as an Interactive Material for Ideation. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–6, 2019.

[278] Y.-C. Lee. Serendipity in scientific discoveries: Some examples in glycosciences. In *The Molecular Immunology of Complex Carbohydrates-3*, pages 3–14. Springer, 2011.

[279] J. Leike, D. Krueger, T. Everitt, M. Martic, V. Maini, and S. Legg. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*, 2018.

[280] J. Leike, M. Martic, V. Krakovna, P. A. Ortega, T. Everitt, A. Lefrancq, L. Orseau, and S. Legg. AI safety gridworlds. *arXiv preprint arXiv:1711.09883*, 2017.

[281] H. M. Lewis and K. N. Laland. Transmission fidelity is the key to the build-up of cumulative culture. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1599):2171–2180, 2012.

[282] S. Li, J. Zhang, P. Li, Y. Wang, and Q. Wang. Influencing factors of driving decision-making under the moral dilemma. *IEEE Access*, 2019.

[283] B. Liang, H. Li, M. Su, P. Bian, X. Li, and W. Shi. Deep Text classification Can be Fooled. *arXiv preprint arXiv:1704.08006*, 2017.

[284] H. Lin and O. Vartanian. A neuroeconomic framework for creative cognition. *Perspectives on Psychological Science*, 13(6):655–677, 2018.

[285] B. Little. What Type of Criminal Are You? 19th-Century Doctors Claimed to Know by Your Face. `https://www.history.com/news/born-criminal-theory-criminology`, 2019. History; accessed 23-May-2020.

[286] P. Liu, Y. Du, and Z. Xu. Machines versus humans: People's biased responses to traffic accidents involving self-driving vehicles. *Accident Analysis & Prevention*, 125:232–240, 2019.

[287] S. Llewellyn. Dream to predict? REM dreaming as prospective coding. *Frontiers in psychology*, 6:1961, 2016.

[288] L.-D. Lord, P. Expert, S. Atasoy, L. Roseman, K. Rapuano, R. Lambiotte, D. J. Nutt, G. Deco, R. L. Carhart-Harris, M. L. Kringelbach, et al. Dynamical exploration of the repertoire of brain networks at rest is modulated by psilocybin. *NeuroImage*, 199:127–142, 2019.

[289] C. Lucchiari and M. E. Vanutelli. Promoting creativity through transcranial direct current stimulation (tDCS). A critical review. *Frontiers in behavioral neuroscience*, 12:167, 2018.

[290] S. Lyubomirsky. Why are some people happier than others? The role of cognitive and motivational processes in well-being. *American psychologist*, 56(3):239, 2001.

[291] P. D. MacLean. *The triune brain in evolution: Role in paleocerebral functions.* Springer Science & Business Media, 1990.

[292] L. Magnani, C. Casadio, and Magnani. *Model-based reasoning in science and technology.* Springer, 2016.

[293] F. A. Mansouri, E. Koechlin, M. G. Rosa, and M. J. Buckley. Managing competing goals—a key role for the frontopolar cortex. *Nature Reviews Neuroscience*, 18(11):645, 2017.

[294] G. Marcus. The next decade in AI: four steps towards robust artificial intelligence. *arXiv preprint arXiv:2002.06177*, 2020.

[295] D. M. Markowitz, R. Laha, B. P. Perone, R. D. Pea, and J. N. Bailenson. Immersive virtual reality field trips facilitate learning about climate change. *Frontiers in Psychology*, 9:2364, 2018.

[296] S. Martin. AI Observatory opens in Berlin on 3 March 2020 . `https://www.itspmagazine.com/itsp-chronicles/ai-village-what-is-ai-safety-and-how-can-we-embrace-and-prepare-for-adversarial-ai`, 2018. ITSP Magazine; accessed 25-April-2020.

[297] A. H. Maslow. *The farther reaches of human nature.* Viking, 1971.

[298] J. Masthoff. Group recommender systems: Combining individual models. In *Recommender systems handbook*, pages 677–702. Springer, 2011.

[299] H. R. Maturana and F. J. Varela. *The tree of knowledge: The biological roots of human understanding.* New Science Library/Shambhala Publications, 1987.

[300] T. Metzinger. Towards a global artificial intelligence charter. *Should we fear artificial intelligence*, pages 27–33, 2018.

[301] L. Meuhlhauser and L. Helm. Intelligence Explosion and Machine Ethics. *Singularity Hypotheses: A Scientific and Philosophical Assessment*, pages 101–126, 2012.

[302] M. Miller and A. Clark. Happily entangled: prediction, emotion, and the embodied mind. *Synthese*, 195(6):2559–2575, 2018.

[303] D. E. Milton. On the ontological status of autism: the 'double empathy problem'. *Disability & Society*, 27(6):883–887, 2012.

[304] J. Mirkovic, P. Reiher, C. Papadopoulos, A. Hussain, M. Shepard, M. Berg, and R. Jung. Testing a collaborative DDoS defense in a red team/blue team exercise. *IEEE Transactions on Computers*, 57(8):1098–1112, 2008.

[305] R. Mirski, M. H. Bickhard, D. Eck, and A. Gut. Encultured minds, not error reduction minds. *Behavioral and Brain Sciences*, 43, 2020.

[306] D. Mobbs, R. Adolphs, M. S. Fanselow, L. F. Barrett, J. E. LeDoux, K. Ressler, and K. M. Tye. Viewpoints: Approaches to defining and investigating fear. *Nature neuroscience*, 22(8):1205–1216, 2019.

[307] A. Month et al. Artful mathematics: The heritage of MC Escher. *Notices of the AMS*, 50(4):446–451, 2003.

[308] J. H. Moor. The nature, importance, and difficulty of machine ethics. *IEEE intelligent systems*, 21(4):18–21, 2006.

[309] C. K. Morewedge, H. Yoon, I. Scopelliti, C. W. Symborski, J. H. Korris, and K. S. Kassam. Debiasing decisions: Improved decision making with a single training intervention. *Policy Insights from the Behavioral and Brain Sciences*, 2(1):129–140, 2015.

[310] J. Mossbridge, B. Goertzel, R. Mayet, E. Monroe, G. Nehat, D. Hanson, and G. Yu. Emotionally-sensitive AI-driven android interactions improve social welfare through helping people access self-transcendent states. vol. In *AI for Social Good Workshop at Neural Information Processing Systems 2018 Conference*, 2018.

[311] L. Muehlhauser and A. Salamon. Intelligence explosion: Evidence and import. In *Singularity hypotheses*, pages 15–42. Springer, 2012.

[312] V. C. Müller and N. Bostrom. Future progress in artificial intelligence: A survey of expert opinion. In *Fundamental issues of artificial intelligence*, pages 555–572. Springer, 2016.

[313] B. Nassi, D. Nassi, R. Ben-Netanel, Y. Mirsky, O. Drokin, and Y. Elovici. Phantom of the ADAS: Phantom Attacks on Driver-Assistance Systems. *IACR Cryptology ePrint Archive*, 2020:85, 2020.

[314] U. Nations. Tier classification for global SDG indicators. 2019.

[315] P. Neekhara, S. Hussain, M. Jere, F. Koushanfar, and J. McAuley. Adversarial Deepfakes: Evaluating Vulnerability of Deepfake Detectors to Adversarial Examples. *arXiv preprint arXiv:2002.12749*, 2020.

[316] R. L. Newman and J. F. Connolly. Determining the role of phonology in silent reading using event-related brain potentials. *Cognitive Brain Research*, 21(1):94–105, 2004.

[317] R. L. Newman and M. F. Joanisse. Modulation of brain regions involved in word recognition by homophonous stimuli: an fMRI study. *Brain Research*, 1367:250–264, 2011.

[318] E. Nivel, K. R. Thórisson, B. R. Steunebrink, H. Dindo, G. Pezzulo, M. Rodríguez, C. Hernández, D. Ognibene, J. Schmidhuber, R. Sanz, et al. Bounded recursive self-improvement. *arXiv preprint arXiv:1312.6764*, 2013.

[319] V. Noreika, J. M. Windt, M. Kern, K. Valli, T. Salonen, R. Parkkola, A. Revonsuo, A. A. Karim, T. Ball, and B. Lenggenhager. Modulating dream experience: Non-invasive brain stimulation over the sensorimotor cortex reduces dream movement. *Scientific Reports*, 10(1):1–19, 2020.

[320] J. Norman and Y. Bar-Yam. Special Operations Forces: A Global Immune System? In *International Conference on Complex Systems*, pages 486–498. Springer, 2018.

[321] S. Nyholm. The ethics of crashes with self-driving cars: A roadmap, I. *Philosophy Compass*, 13(7):e12507, 2018.

[322] S. Nyholm and J. Smids. The ethics of accident-algorithms for self-driving cars: An applied trolley problem? *Ethical theory and moral practice*, 19(5):1275–1289, 2016.

[323] OECD.AI. OECD AI Policy Observatory . `https://oecd.ai/`, 2020. Online; accessed 25-April-2020.

[324] E. Ok. Elements of order theory. *Unpublished book, New York University.[411]*, 2011.

[325] S. Oosterwijk, K. A. Lindquist, E. Anderson, R. Dautoff, Y. Moriguchi, and L. F. Barrett. States of mind: Emotions, body feelings, and thoughts share distributed neural networks. *NeuroImage*, 62(3):2110–2128, 2012.

[326] N. Papernot, P. McDaniel, A. Sinha, and M. Wellman. Towards the science of security and privacy in machine learning. *arXiv preprint arXiv:1611.03814*, 2016.

[327] N. Papernot, P. McDaniel, A. Sinha, and M. Wellman. Towards the science of security and privacy in machine learning. *arXiv preprint arXiv:1611.03814*, 2016.

[328] D. Parfit. *Reasons and persons.* OUP Oxford, 1984.

[329] B. Paris and J. Donovan. Deepfakes and Cheap Fakes. *United States of America: Data & Society*, 2019.

[330] T. Parr, L. Da Costa, and K. Friston. Markov blankets, information geometry and stochastic thermodynamics. *Philosophical Transactions of the Royal Society A*, 378(2164):20190159, 2019.

[331] T. Partala and V. Surakka. The effects of affective interventions in human–computer interaction. *Interacting with computers*, 16(2):295–309, 2004.

[332] L. Pascu. Biometric software that allegedly predicts criminals based on their face sparks industry controversy. `https://www.biometricupdate.com/202005/biometric-software-that-allegedly-predicts-criminals-based-on-their-face-sparks-industry-controversy`, 2020. Biometric; accessed 23-May-2020.

[333] A. R. Pathak, S. Bhalsing, S. Desai, M. Gandhi, and P. Patwardhan. Deep learning model for facial emotion recognition. In *Proceedings of ICETIT 2019*, pages 543–558. Springer, 2020.

[334] C. Pattamadilok, V. Chanoine, C. Pallier, J.-L. Anton, B. Nazarian, P. Belin, and J. C. Ziegler. Automaticity of phonological and semantic processing during visual word recognition. *NeuroImage*, 149:244–255, 2017.

[335] K. Perrykkad and J. Hohwy. Modelling Me, Modelling You: the Autistic Self. *Review Journal of Autism and Developmental Disorders*, pages 1–31, 2019.

[336] D. Peters, K. Vold, D. Robinson, and R. A. Calvo. Responsible AI—Two Frameworks for Ethical Design Practice. *IEEE Transactions on Technology and Society*, 1(1):34–47, 2020.

[337] C. Peterson. *A primer in positive psychology.* Oxford University Press, 2006.

[338] F. Pistono and R. V. Yampolskiy. Unethical research: how to create a malevolent artificial intelligence. *arXiv preprint arXiv:1605.02817*, 2016.

[339] K. Popper. R.(1957). The Poverty of Historicism, 1957.

[340] K. Popper. In P. A. Schilpp, editor, *The Philosophy of Karl Popper*, volume 2, page 1015. Open Court Press, 1974.

[341] K. Popper. *Conjectures and refutations: The growth of scientific knowledge.* routledge, 2014.

[342] A. Potapov. Technological Singularity: What Do We Really Know? *Information*, 9(4):82, 2018.

[343] E. Pranav, S. Kamal, C. S. Chandran, and M. Supriya. Facial Emotion Recognition Using Deep Convolutional Neural Network. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pages 317–320. IEEE, 2020.

[344] J. Pujol, I. Batalla, O. Contreras-Rodríguez, B. J. Harrison, V. Pera, R. Hernández-Ribas, E. Real, L. Bosa, C. Soriano-Mas, J. Deus, et al. Breakdown in the brain network subserving moral judgment in criminal psychopathy. *Social cognitive and affective neuroscience*, 7(8):917–923, 2011.

[345] D. Quesnel and B. E. Riecke. Are You Awed Yet? How Virtual Reality Gives Us Awe and Goose Bumps. *Frontiers in Psychology*, 9, 2018.

[346] A. Radford, J. Wu, D. Amodei, D. Amodei, J. Clark, M. Brundage, and I. Sutskever. Better language models and their implications. *OpenAI Blog https://openai.com/blog/better-language-models*, 2019.

[347] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019.

[348] R. Rafal and A. Kenji. 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.

[349] J. Rajendran, V. Jyothi, and R. Karri. Blue team red team approach to hardware trust assessment. In *2011 IEEE 29th international conference on computer design (ICCD)*, pages 285–288. IEEE, 2011.

[350] A. Recupero, S. Triberti, C. Modesti, and A. Talamo. Mixed Reality for Cross-Cultural Integration: Using Positive Technology to Share Experiences and Promote Communication. *Frontiers in psychology*, 9:1223, 2018.

[351] A. Rege. Incorporating the human element in anticipatory and dynamic cyber defense. In *2016 IEEE International Conference on Cybercrime and Computer Forensic (ICCCF)*, pages 1–7. IEEE, 2016.

[352] A. Rege, Z. Obradovic, N. Asadi, B. Singer, and N. Masceri. A temporal assessment of cyber intrusion chains using multidisciplinary frameworks and methodologies. In *2017 International Conference on Cyber Situational Awareness, Data Analytics And Assessment (Cyber SA)*, pages 1–7. IEEE, 2017.

[353] A. Renda. Ethics, algorithms and self-driving cars–a CSI of the 'trolley problem'. *CEPS Policy Insight*, (2018/02), 2018.

[354] M. Reuter and C. Montag. *Neuroeconomics*. Springer, 2016.

[355] E. Rietveld. The affordances of art for making technologies. 2019.

[356] A. Riikonen. Decide, Disrupt, Destroy: Information Systems in Great Power Competition with China. *Strategic Studies Quarterly*, 13(4), 2019.

[357] R. M. Roberts. Serendipity: Accidental discoveries in science. *Serendipity: Accidental Discoveries in Science, by Royston M. Roberts, pp. 288. ISBN 0-471-60203-5. Wiley-VCH, June 1989.*, page 288, 1989.

[358] R. S. Rosenberg, S. L. Baughman, and J. N. Bailenson. Virtual superheroes: Using superpowers in virtual reality to encourage prosocial behavior. *PloS one*, 8(1):e55003, 2013.

[359] A. Rotsidis, A. Theodorou, J. J. Bryson, and R. H. Wortham. Improving robot transparency: An investigation with mobile augmented reality. In *28th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), New Delhi. IEEE*, 2019.

[360] D. Rudrauf, D. Bennequin, I. Granic, G. Landini, K. Friston, and K. Williford. A mathematical model of embodied consciousness. *Journal of theoretical biology*, 428:106–131, 2017.

[361] D. Rudrauf, D. Bennequin, and K. Williford. The Moon Illusion explained by the Projective Consciousness Model. *Journal of Theoretical Biology*, page 110455, 2020.

[362] M. A. Runco. *Critical creative processes.* Hampton Press, 2003.

[363] S. Russel, P. Norvig, et al. *Artificial intelligence: a modern approach.* Pearson Education Limited, 2013.

[364] S. Russell. How to Stop Superhuman A.I. Before It Stops Us. `https://www.nytimes.com/2019/10/08/opinion/artificial-intelligence.html?module=inline`. Accessed: 2019-11-21.

[365] S. Russell, D. Dewey, and M. Tegmark. Research priorities for robust and beneficial artificial intelligence. *AI Magazine*, 36(4):105–114, 2015.

[366] N. Sajid, P. J. Ball, and K. J. Friston. Active inference: demystified and compared. *arXiv: Artificial Intelligence*, 2019.

[367] C. Sancaktar and P. Lanillos. End-to-end pixel-based deep active inference for body perception and action. *arXiv preprint arXiv:2001.05847*, 2019.

[368] M. Sánchez-Francisco, P. Díaz, F. Fabiano, and I. Aedo. Engaging Users with an AR Pervasive Game for Personal Urban Awareness. In *Proceedings of the XX International Conference on Human Computer Interaction*, page 6. ACM, 2019.

[369] T. Sato, J. Shen, N. Wang, Y. J. Jia, X. Lin, and Q. A. Chen. Security of Deep Learning based Lane Keeping System under Physical-World Adversarial Attack. *arXiv preprint arXiv:2003.01782*, 2020.

[370] S. Savage-Rumbaugh, I. Roffman, S. Lingomo, and E. Pugh. The fully conscious ape. *International Journal of Comparative Psychology*, 31, 2018.

[371] C. Schein and K. 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.

[372] C. Schein and K. Gray. Moralization and harmification: The dyadic loop explains how the innocuous becomes harmful and wrong. *Psychological Inquiry*, 27(1):62–65, 2016.

[373] C. Schein and K. Gray. The theory of dyadic morality: Reinventing moral judgment by redefining harm. *Personality and Social Psychology Review*, 22(1):32–70, 2018.

[374] C. Schein, N. Hester, and K. 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.

[375] J. Schmidhuber. Gödel machines: Fully self-referential optimal universal self-improvers. In *Artificial general intelligence*, pages 199–226. Springer, 2007.

[376] M. Schroeder. Teleology, agent-relative value, and 'good'. *Ethics*, 117(2):265–295, 2007.

[377] J. Schulkin and P. Sterling. Allostasis: A brain-centered, predictive mode of physiological regulation. *Trends in neurosciences*, 2019.

[378] A. Schüren. Vom Suchen und Finden. *MEDIENwissenschaft: Rezensionen— Reviews*, 31(2-3):163, 2014.

[379] N. S. Schutte and E. J. Stilinović. Facilitating empathy through virtual reality. *Motivation and emotion*, 41(6):708–712, 2017.

[380] P. J. Scott and R. V. Yampolskiy. Classification Schemas for Artificial Intelligence Failures. *arXiv preprint arXiv:1907.07771*, 2019.

[381] M. S. Seidenberg and J. L. McClelland. A distributed, developmental model of word recognition and naming. *Psychological review*, 96(4):523, 1989.

[382] M. E. Seligman. *Flourish: A visionary new understanding of happiness and well-being*. Simon and Schuster, 2012.

[383] M. E. Seligman and M. Csikszentmihalyi. Positive psychology: An introduction. In *Flow and the foundations of positive psychology*, pages 279–298. Springer, 2014.

[384] A. K. Seth, W. Wiese, T. Metzinger, and J. M. Windt. Inference to the best prediction. *Open MIND. MIND Group, Frankfurt am Main. Shew, WL, Yang, H., Yu, S., Roy, R., Plenz, D*, pages 55–63, 2011.

[385] G. Sevinc, H. Gurvit, and R. N. Spreng. Salience network engagement with the detection of morally laden information. *Social cognitive and affective neuroscience*, 12(7):1118–1127, 2017.

[386] O. Sezer, F. Gino, and M. H. Bazerman. Ethical blind spots: Explaining unintentional unethical behavior. *Current Opinion in Psychology*, 6:77–81, 2015.

[387] L. Shi, R. E. Beaty, Q. Chen, J. Sun, D. Wei, W. Yang, and J. Qiu. Brain Entropy is Associated with Divergent Thinking. *Cerebral Cortex*, 2019.

[388] Z. Shi, G. Ma, S. Wang, and J. Li. Brain-machine collaboration for cyborg intelligence. In *International Conference on Intelligent Information Processing*, pages 256–266. Springer, 2016.

[389] C. J. Simon. Ethics and Artificial General Intelligence: Technological Prediction as a Groundwork for Guidelines. In *2019 IEEE International Symposium on Technology and Society (ISTAS)*, pages 1–6. IEEE, 2019.

[390] T. Simonite. “When it comes to Gorillas, Google Photos Remains Blind”. `https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/`, 2018. Online; accessed 17-October-2019.

[391] D. K. Simonton. Creative thought as blind variation and selective retention: Why creativity is inversely related to sightedness. *Journal of Theoretical and Philosophical Psychology*, 33(4):253, 2013.

[392] N. Soares and B. Fallenstein. Agent foundations for aligning machine intelligence with human interests: a technical research agenda. In *The Technological Singularity*, pages 103–125. Springer, 2017.

[393] A. Sorensen. The uncertain origins of fire-making by humans: The state of the art and smouldering questions. *Mitteilungen der Gesellschaft für Urgeschichte*, 28:11–50, 12 2019.

[394] K. Sotala and R. V. Yampolskiy. Responses to catastrophic AGI risk: a survey. *Physica Scripta*, 90(1):018001, 2014.

[395] E. R. Stepanova, D. Quesnel, and B. Riecke. Transformative experiences become more accessible through virtual reality. In *2018 IEEE Workshop on Augmented and Virtual Realities for Good (VAR4Good)*, pages 1–3. IEEE, 2018.

[396] C. Stupp. Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case . https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402, 2019. Wall Street Journal; accessed 24-May-2020.

[397] J. Sun, Q. Chen, Q. Zhang, Y. Li, H. Li, D. Wei, W. Yang, and J. Qiu. Training your brain to be more creative: brain functional and structural changes induced by divergent thinking training. *Human brain mapping*, 37(10):3375–3387, 2016.

[398] L. R. Sütfeld, B. V. Ehinger, P. König, and G. Pipa. How does the method change what we measure? Comparing virtual reality and text-based surveys for the assessment of moral decisions in traffic dilemmas. *PsyArXiv*, 2019.

[399] L. R. Sütfeld, R. Gast, P. König, and G. Pipa. Using virtual reality to assess ethical decisions in road traffic scenarios: applicability of value-of-life-based models and influences of time pressure. *Frontiers in behavioral neuroscience*, 11:122, 2017.

[400] J. Taylor, E. Yudkowsky, P. LaVictoire, and A. Critch. Alignment for advanced machine learning systems. *Machine Intelligence Research Institute*, 2016.

[401] M. Tegmark. *Life 3.0: Being human in the age of artificial intelligence*. Knopf, 2017.

[402] Tencent Keen Security Lab. Experimental Security Research of Tesla Autopilot . https://keenlab.tencent.com/en/whitepapers/Experimental_Security_Research_of_Tesla_Autopilot.pdf, 2019. Online; accessed 25-April-2020.

[403] The Agency for Digital Italy. Italian Observatory on Artificial Intelligence . https://ia.italia.it/en/ai-observatory/, 2020. Online; accessed 25-April-2020.

[404] K. R. Thórisson. A new constructivist AI: from manual methods to self-constructive systems. In *Theoretical Foundations of Artificial General Intelligence*, pages 145–171. Springer, 2012.

[405] K. R. Thórisson, J. Bieger, X. Li, and P. Wang. Cumulative learning. In *International Conference on Artificial General Intelligence*, pages 198–208. Springer, 2019.

[406] M. Tomasello. The role of roles in uniquely human cognition and sociality. *Journal for the Theory of Social Behaviour*, 50(1):2–19, 2020.

[407] R. Tomsett, A. Widdicombe, T. Xing, S. Chakraborty, S. Julier, P. Gurram, R. Rao, and M. Srivastava. Why the Failure? How Adversarial Examples Can Provide Insights for Interpretable Machine Learning. In *2018 21st International Conference on Information Fusion (FUSION)*, pages 838–845. IEEE, 2018.

[408] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017.

[409] D. Trott. Deceiving Machines: Sabotaging Machine Learning. *CHANCE*, 33(2):20–24, 2020.

[410] J. Tsao, C. Ting, and C. Johnson. Creative outcome as implausible utility. *Review of General Psychology*, 23(3):279–292, 2019.

[411] A. Turchin, D. Denkenberger, and B. Green. Global Solutions vs. Local Solutions for the AI Safety Problem. *Big Data and Cognitive Computing*, 3(1):16, Feb 2019.

[412] J. Uesato, B. O'Donoghue, A. v. d. Oord, and P. Kohli. Adversarial risk and the dangers of evaluating against weak attacks. *arXiv preprint arXiv:1802.05666*, 2018.

[413] J. Uher. Human uniqueness explored from the uniquely human perspective: Epistemological and methodological challenges. *Journal for the Theory of Social Behaviour*, 2020.

[414] I. Van de Poel. Translating values into design requirements. In *Philosophy and engineering: Reflections on practice, principles and process*, pages 253–266. Springer, 2013.

[415] O. van der Himst and P. Lanillos. Deep Active Inference for Partially Observable MDPs. *arXiv preprint arXiv:2009.03622*, 2020.

[416] E. van Foeken, L. J. Kester, and M. van Iersel. Real-time common awareness in communication constrained sensor systems. In *2009 12th International Conference on Information Fusion*, pages 118–125. IEEE, 2009.

[417] A. van Loon, J. Bailenson, J. Zaki, J. Bostick, and R. Willer. Virtual reality perspective-taking increases cognitive empathy for specific others. *PloS one*, 13(8):e0202442, 2018.

[418] C. P. van Schaik, G. R. Pradhan, and C. Tennie. Teaching and curiosity: sequential drivers of cumulative cultural evolution in the hominin lineage. *Behavioral ecology and sociobiology*, 73(1):2, 2019.

[419] L. R. Varshney, F. Pinel, K. Varshney, D. Bhattacharjya, A. Schörgendorfer, and Y.-M. Chee. A big data approach to computational creativity: The curious case of chef watson. *IBM Journal of Research and Development*, 63(1):7–1, 2019.

[420] S. P. Veissière, A. Constant, M. J. Ramstead, K. J. Friston, and L. J. Kirmayer. TTOM in Action: Refining the Variational Approach to Cognition and Culture Short title. *Behavioral and Brain Sciences*, in press, 2020.

[421] B. Verbeek. Consequentialism, rationality and the relevant description of outcomes. *Economics & Philosophy*, 17(2):181–205, 2001.

[422] R. Vinuesa, H. Azizpour, I. Leite, M. Balaam, V. Dignum, S. Domisch, A. Felländer, S. D. Langhans, M. Tegmark, and F. F. Nerini. The role of artificial intelligence in achieving the sustainable development goals. *Nature Communications*, 11(1):1–10, 2020.

[423] E. Wallace, S. Feng, N. Kandpal, M. Gardner, and S. Singh. Universal adversarial triggers for NLP. *arXiv preprint arXiv:1908.07125*, 2019.

[424] P. Wang. Motivation management in AGI systems. In *International Conference on Artificial General Intelligence*, pages 352–361. Springer, 2012.

[425] P. Wang, X. Li, and P. Hammer. Self in NARS, an AGI System. *Frontiers in Robotics and AI*, 5:20, 2018.

[426] W. Warner and J. Hirschberg. Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media*, pages 19–26. Association for Computational Linguistics, 2012.

[427] M. Way. What I cannot create, I do not understand. *Journal of Cell Science*, 130:2941–2942, 2017.

[428] P. Werkhoven, L. Kester, and M. Neerincx. Telling autonomous systems what to do. In *Proceedings of the 36th European Conference on Cognitive Ergonomics*, pages 1–8, 2018.

[429] K. L. Wheat, P. L. Cornelissen, S. J. Frost, and P. C. Hansen. During visual word recognition, phonology is accessed within 100 ms and may be mediated by a speech production code: evidence from magnetoencephalography. *Journal of Neuroscience*, 30(15):5229–5233, 2010.

[430] W. Wiese. *Perceptual Presence in the Kuhnian-Popperian Bayesian Brain: A Commentary on Anil K. Seth.* Johannes Gutenberg-Universität Mainz, 2016.

[431] D. E. Wildman, M. Uddin, G. Liu, L. I. Grossman, and M. Goodman. Implications of natural selection in shaping 99.4% nonsynonymous DNA identity between humans and chimpanzees: enlarging genus Homo. *Proceedings of the national Academy of Sciences*, 100(12):7181–7188, 2003.

[432] K. Williford, D. Bennequin, K. Friston, and D. Rudrauf. The projective consciousness model and phenomenal selfhood. *Frontiers in psychology*, 9:2571, 2018.

[433] B. Wilson, J. Hoffman, and J. Morgenstern. Predictive inequity in object detection. *arXiv preprint arXiv:1902.11097*, 2019.

[434] H. Wilson, A. Theodorou, and J. J. Bryson. Slam the Brakes: Perceptions of Moral Decisions in Driving Dilemmas. In *International Workshop in Artificial Intelligence Safety (AISafety), IJCAI, Macau*, 2019.

[435] L. B. Wilson, J. R. Tregellas, E. Slason, B. E. Pasko, and D. C. Rojas. Implicit phonological priming during visual word recognition. *Neuroimage*, 55(2):724–731, 2011.

[436] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv*, abs/1910.03771, 2019.

[437] J. Wolfe and J. Dastin. U.S. government study finds racial bias in facial recognition tools . `https://www.reuters.com/article/us-usa-crime-face/u-s-government-study-finds-racial-bias-in-facial-recognition-tools-idUSKBN1YN2V1`, 2020. Reuters; accessed 23-May-2020.

[438] C. Wong. Dancin seq2seq: Fooling text classifiers with adversarial text example generation. *arXiv preprint arXiv:1712.05419*, 2017.

[439] E. Y. Wong, T. Kwong, and M. Pegrum. Learning on mobile augmented reality trails of integrity and ethics. *Research and Practice in Technology Enhanced Learning*, 13(1):22, 2018.

[440] X. Wu and X. Zhang. Automated inference on criminality using face images. *arXiv preprint arXiv:1611.04135*, pages 4038–4052, 2016.

[441] R. Yampolskiy and J. Fox. Safety engineering for artificial general intelligence. *Topoi*, 32(2):217–226, 2013.

[442] R. V. Yampolskiy. Leakproofing Singularity-Artificial Intelligence Confinement Problem. *Journal of Consciousness Studies JCS*, 2012.

[443] R. V. Yampolskiy. Artificial intelligence safety engineering: Why machine ethics is a wrong approach. In *Philosophy and theory of artificial intelligence*, pages 389–396. Springer, 2013.

[444] R. V. Yampolskiy. Utility function security in artificially intelligent agents. *Journal of Experimental & Theoretical Artificial Intelligence*, 26(3):373–389, 2014.

[445] R. V. Yampolskiy. *Artificial Superintelligence: A Futuristic Approach.* Chapman and Hall/CRC, 2015.

[446] R. V. Yampolskiy. Taxonomy of pathways to dangerous artificial intelligence. In *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[447] R. V. Yampolskiy. Detecting qualia in natural and artificial agents. *arXiv preprint arXiv:1712.04020*, 2017.

[448] R. V. Yampolskiy. *Artificial Intelligence Safety and Security.* Chapman and Hall/CRC, 2018.

[449] R. V. Yampolskiy. Personal Universes: A Solution to the Multi-Agent Value Alignment Problem. *arXiv preprint arXiv:1901.01851*, 2019.

[450] R. V. Yampolskiy. Predicting future AI failures from historic examples. *foresight*, 2019.

[451] R. V. Yampolskiy and M. Spellchecker. Artificial intelligence safety and cybersecurity: A timeline of AI failures. *arXiv preprint arXiv:1610.07997*, 2016.

[452] N. Yee and J. N. Bailenson. Walk a mile in digital shoes: The impact of embodied perspective-taking on the reduction of negative stereotyping in immersive virtual environments. *Proceedings of PRESENCE*, 24:26, 2006.

[453] E. Yudkowsky. Creating friendly AI 1.0: The analysis and design of benevolent goal architectures. *The Singularity Institute, San Francisco, USA*, 2001.

[454] E. Yudkowsky. Coherent extrapolated volition. *Singularity Institute for Artificial Intelligence*, 2004.

[455] E. Yudkowsky. Cognitive biases potentially affecting judgment of global risks. *Global catastrophic risks*, 1(86):13, 2008.

[456] E. Yudkowsky. The AI alignment problem: why it is hard, and where to start. *Symbolic Systems Distinguished Speaker*, 2016.

[457] E. Yudkowsky, A. Salamon, C. Shulman, S. Kaas, T. McCabe, R. Nelson, et al. Reducing long-term catastrophic risks from artificial intelligence. *The Singularity Institute, San Francisco*, 2010.

[458] K. Zetter. Researchers Easily Trick Cylance's AI-Based Antivirus Into Thinking Malware Is 'Goodware' . `https://www.vice.com/en_us/article/9kxp83/researchers-easily-trick-cylances-ai-based-antivirus-into-thinking-malware-is-goodware`, 2019. Motherboard Tech by Vice; accessed 24-May-2020.

[459] H. Zhou, M. Huang, T. Zhang, X. Zhu, and B. Liu. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[460] S. Ziesche. Innovative Big Data Approaches for Capturing and Analyzing Data to Monitor and Achieve the SDGs. *Report of the United Nations Economic and Social*

Commission for Asia and the Pacific: Subregional Office for East and North-East Asia (ESCAP-ENEA), 2017.

[461] S. Ziesche. Potential Synergies Between The United Nations Sustainable Development Goals And The Value Loading Problem In Artificial Intelligence. *Maldives National Journal of Research*, 6:47, 06 2018.

[462] S. Ziesche. AI & Global Governance: A Seat at the Negotiating Table for AI? Opportunities and Risks. `https://cpr.unu.edu/a-seat-at-the-negotiating-table.html`, 2019. United Nations University; accessed 17-October-2019.

[463] D. Zuromski, A. Fedyniuk, and E. Maria. Can New Technologies Make Us More Human? An Inquiry on VR Technologies in Social Cognition. *Frontiers in psychology*, 9:705, 2018.

Curriculum Vitae

Nadisha-Marie Aliman

Homepage: https://nadishamarie.jimdo.com/
Email: nadishamarie.aliman@gmail.com

Education

03/2019 – 12/2020 **Utrecht University**

PhD degree in Computer Science. Anticipated defense date: December 2, 2020

(in approx. 1,5 instead of 4 years standard period of study)

- PhD Thesis:
  „Hybrid Cognitive-Affective Strategies for AI Safety”
- Specializations:
  AI Safety and Security, Cognitive Science, Affective Science and Affective Computing, Adversarial AI, Creativity, Meaningful Control of Intelligent Systems

04/2017 – 09/2018 **University of Stuttgart**

Master’s degree in Computational Linguistics - summa cum laude
(in 3 instead of 4 semesters standard period of study)

- Master Thesis:
  „Cognitive Defense Mechanisms against Adversarial Examples in Natural Language Processing”
- Specializations:
  Applied Natural Language Processing, Cognitive Science, Adversarial Machine Learning, Machine Ethics, Emotion and Sentiment Analysis, Affective Computing

10/2014 – 03/2017 **Saarland University**

Bachelor’s degree in Computational Linguistics
(in 5 instead of 6 semesters standard period of study)

- Bachelor Thesis:
  „Ladder Networks for Named Entity Recognition”
- Supplementary subject: Computer Science
- Specializations:
  Deep Learning, Sentiment Analysis, Information Extraction, Security and Privacy

10/2009 – 06/2014 **University of Stuttgart and FH Kaiserslautern**

Orientation phase, modules out of diverse courses of study (i.a. „History of Science and Technology“)

09/2006 – 06/2009 **Luisengymnasium Düsseldorf**

Abitur (in total shorter education through the skipping of 3 grades at elementary school)