https://www.wageningenacademic.com/doi/pdf/10.3920/978-90-8686-922-0_4 - Tuesday, October 04, 2022 6:24:32 AM - Utrecht University IP Address:131.211.12.11

4. Moral programming

Crafting a flexible heuristic moral meta-model for meaningful AI control in pluralistic societies

Nadisha-Marie Aliman[1] and Leon Kester[2*]

[1]Utrecht University, Faculty of Science, Department of Information and Computing Sciences, Princetonplein 5, 3584 CC Utrecht, the Netherlands; [2]TNO Netherlands, Intelligent Autonomous Systems, Postbus 96864, 2509 JG The Hague, the Netherlands; leon.kester@tno.nl

Abstract

Artificial Intelligence (AI) permeates more and more application domains. Its progress regarding scale, speed, and scope magnifies potential societal benefits but also ethically and safety relevant risks. Hence, it becomes vital to seek a meaningful control of present-day AI systems (i.e. tools). For this purpose, one can aim at counterbalancing the increasing problem-solving ability of AI with boundary conditions core to human morality. However, a major problem is that morality exists in a context-sensitive steadily shifting explanatory sphere co-created by humans using natural language – which is inherently ambiguous at multiple levels and neither machine-understandable nor machine-readable. A related problem is what we call epistemic dizziness, a phenomenon linked to the inevitable circumstance that one could always be wrong. Yet, while universal doubt cannot be eliminated from morality, it need not be magnified if the potential/requirement for steady refinements is anticipated by design. Thereby, morality pertains to the set of norms and values enacted at the level of a society, other not nearer specified collectives of persons, or at the level of an individual. Norms are instrumental in attaining the fulfilment of values, the latter being an umbrella term for all that seems decisive for distinctions between right and wrong – a central object of study in ethics. In short, for a meaningful control of AI against the background of the changing context-sensitive and linguistically moulded nature of human morality, it is helpful to craft descriptive and thus sufficiently flexible AI-readable heuristic models of morality. In this way, the problem-solving ability of AI could be efficiently funnelled through these updatable models so as to ideally boost the benefits and mitigate the risks at the AI deployment stage with the conceivable side-effect of improving human moral conjectures. For this purpose, we introduced

Bart Wernaart (ed.)
Moral design and technology
DOI: 10.3920/978-90-8686-922-0_4, © Wageningen Academic Publishers 2022

https://www.wageningenacademic.com/doi/pdf/10.3920/978-90-8686-922-0_4 - Tuesday, October 04, 2022 6:24:32 AM - Utrecht University IP Address:131.211.12.11

a novel transdisciplinary framework denoted augmented utilitarianism (AU) (Aliman and Kester, 2019b), which is formulated from a meta-ethical stance. AU attempts to support the human-centred task to harness human norms and values to explicitly and traceably steer AI before humans themselves get unwittingly and unintelligibly steered by the obscurity of AI's deployment. Importantly, AU is descriptive, non-normative, and explanatory (Aliman, 2020), and is not to be confused with normative utilitarianism. (While normative ethics pertains to 'what one ought to do', descriptive ethics relates to empirical studies on human ethical decision-making.) This chapter offers the reader a compact overview of how AU coalesces elements from AI, moral psychology, cognitive and affective science, mathematics, systems engineering, cybernetics, and epistemology to craft a generic scaffold able to heuristically encode given moral frameworks in a machine-readable form. We thematise novel insights and also caveats linked to advanced AI risks yielding incentives for future work.

Key concepts

- For meaningful AI control, one needs to channel AI's ability using human morality
- Human morality is context-sensitive, shifting, and ambiguously embedded in language
- It is difficult to make human morality machine-readable
- Descriptive ethics is the study of how human morality is enacted and not what one ought to do
- Augmented utilitarianism (AU) is not to be confused with utilitarianism. AU is a transdisciplinary framework for meaningful AI control of a meta-ethical, descriptive, explanatory, and non-normative nature
- AU offers a generic AI-readable scaffold for heuristic moral models facilitating AI control
- Thereby, AU coalesces elements from descriptive ethics, epistemology, and a wide range of scientific disciplines leading to novel insights that are relevant for future work on AI control

https://www.wageningenacademic.com/doi/pdf/10.3920/978-90-8686-922-0_4 - Tuesday, October 04, 2022 6:24:32 AM - Utrecht University IP Address:131.211.12.11

4.1 A theoretical solution for meaningful artificial intelligence control

Under augmented utilitarianism (AU), one can subdivide the core practical task of meaningfully controlling AI via machine-readable heuristic moral models[2] into three challenging subtasks. The first two subtasks require theoretical epistemic and scientific foundations. The third subtask pertains to realising a complex multi-stakeholder implementation. While this section focuses on the first two theoretical subtasks, the latter is presented in the next Section 4.2. Firstly, while it is habitual for humans to mentally extract morally relevant events from the ongoing experience to which moral judgements are applied, AI neither automatically performs such assessments nor the kind of moral extraction procedure required in the first place. One could call the latter a moral chunking problem, i.e. the problem of locating a meaningful abstraction level for heuristic moral models. As an illustrative analogy from a different context, one can consider the example of human language acquisition where infants face the linguistic chunking problem of dividing incoming information, such as, for instance, auditory information, into meaningful chunks (Isbilen and Christiansen, 2020). A similar problem is also encountered by adults in the process of second language acquisition. Without prior semantic knowledge it is often daunting to identify the boundaries of individual words or multi-word expressions from speech. Secondly, even if an AI were programmed with an appropriate moral abstraction level for a selected ethical framework and the parameters of relevance within that chosen framework were filled in, you are faced with an old epistemic: what if you were wrong and changed your mind or what if you were uncertain about your choices? Below, we look at some theoretical solutions to these two AU subtasks.

moral chunking problem

4.1.1 Moral chunking

As mentioned earlier, AU is formulated from a non-normative meta-ethical stance which signifies that instead of discussing what one ought to do (as is the case in utilitarianism), it is instead located at a higher level focusing on what morality itself is, how it is applied and why. However, instead of having philosophical aspirations, AU engages in these deliberations for quintessentially pragmatic safety-oriented socio-technological reasons. To this end, AU employs knowledge from different scientific disciplines to get a clearer view of the nature of human morality, and integrates an engineering-relevant systems

[2] Note that these models are heuristics since one is always only provisionally modelling utterly complex moving targets of morality. As a result, both moral targets and abstract models thereof are fallible and not necessarily optimal at any point in time. In short, it is acknowledged that one could always be wrong.

https://www.wageningenacademic.com/doi/pdf/10.3920/978-90-8686-922-0_4 - Tuesday, October 04, 2022 6:24:32 AM - Utrecht University IP Address:131.211.12.11

perspective to finally work out technically feasible solutions. It is thus in such a pragmatic safety-oriented context that AU studies this first subtask of moral chunking for meaningful AI control. To start with, AU can be conceptually understood as a sort of input-agnostic interface between humans equipped with a certain instance of an ethical framework and the AI they intend to align with this framework. To be able to fulfil this generic interface role, AU must search for efficient moral chunking with the capacity to ideally be applicable to most ethical frameworks that society would select to control AI. In the context of autonomous vehicles, recent work found that 'it is essential to increase the public's confidence that the values of a pluralistic society are accounted for' (Dubljević *et al.*, 2021). This may apply more generally to any AI deployed at societal scales. In this vein, AU targets what one could conceive of as a possible smallest heuristic moral superset (SHMS) capturing the plurality of candidate ethical frameworks available in practice for moral programming – with the pre-condition that AU must flag known formal inconsistencies to forestall predictable practical safety problems in AI deployment. The latter shows why moral chunking in AU cannot merely be descriptive but must be simultaneously explanatory. From a systems engineering perspective, to specifically strive for an SHMS is not only of interest by virtue of being simpler but it is also more energy-efficient since one does not integrate more parameters than necessary (which would e.g. afford more computing power) while still acting on the requirement to integrate moral pluralism phenomena whose existence has been often corroborated in moral psychology studies (Schein and Gray, 2018).

smallest heuristic moral superset

The three classical normative ethical frameworks that are often mentioned as candidates for moral programming are virtue ethics, deontic ethics, and utilitarianism. Note thereby that moral chunking in AU is agnostic to whether the candidate ethical framework is in fact of a normative, relativistic, comparative, purely pragmatic, or other nature. Here, we use the example of the three normative frameworks due to their salience in AI ethics and AI safety debates. While virtue ethics focuses on the virtues exhibited by the agent, deontology foregrounds the action itself and utilitarianism the consequences of the action to the receiver of the action, also called the patient. In moral programming, the agent carrying out the action is often the AI while the patient is a human. However, in human-machine teaming contexts, agents and patients can be either humans or AIs depending on the situation at hand, which means that sometimes a task can in theory also be purposefully limited to AIs (e.g. if the collaborative task were too dangerous to be performed by human agents). In the AI field, the widespread use of utility-maximising AI systems in the form of reinforcement learning agents, modelled in the fashion of utilitarian economics, established utility functions as default standard for moral programming in their characteristic of reflecting supposed rational decision-making. However, there

https://www.wageningenacademic.com/doi/pdf/10.3920/978-90-8686-922-0_4 - Tuesday, October 04, 2022 6:24:32 AM - Utrecht University IP Address:131.211.12.11

can be a difference between: (1) what humans think they ought to do; and (2) what they want an AI to do in practice. Furthermore, there can be a difference between the former and what humans can actually do as well as between the latter and what an AI itself can do in practice. The practical consequences of such caveats (re)appear throughout our analysis. In the following paragraph, we first briefly shed light on key scientific insights that are of relevance for moral chunking under AU.

Recent moral psychology studies suggest that human morality is based on constructions, is quintessentially perceiver-dependent, context-sensitive, multi-faceted, and is linked to diverse often implicit mind perception phenomena and unconscious assessments. For instance, empirical evaluations in the context of the Agents, Deeds, and Consequences (Dubljević *et al.*, 2018) model (ADC-model) corroborated that moral judgements unify intuitions from all three common normative frameworks mentioned earlier, which according to the authors 'provides an explanation for the intuitive appeal of dominant moral theories' (Dubljević *et al.*, 2018). Following the theory of dyadic morality (TDM) (Schein and Gray, 2018), human moral judgements are harm-based constructions that relate to a perceiver-dependent cognitive template encoding how much damage an intentional agent causes to a vulnerable patient. In this way, TDM frames morality as exhibiting unity in the variety of perceived harm. In view of the moral judgment as categorisation (MJAC) framework (McHugh *et al.*, 2021), 'understanding the making of moral judgments requires accounting for the full complexity and variability of our moral judgments'. Namely, MJAC conjectures (in the spirit of constructionist frameworks in psychology (Oosterwijk *et al.*, 2012)) that moral judgements are perceiver-dependent cognitive-affective constructions based on domain-general elements (not limited to harm) – describing morality as one of the potentially infinite set of situated embodied conceptualisations, i.e. categories that human minds can bring forth and learn to bring forth. Generally, constructionist theories in psychology and affective neuroscience reveal that human cognition is inseparably affective (Hoemann and Feldman Barrett, 2019; Kleckner *et al.*, 2017), and so the assumption that humans are cold rational agents acting according to an affect-free mathematical strategy is a myth. As stated by Barrett (2017), the human brain is anatomically structured in such a way that no action is free of interception and affect.

Overall, one can conclude that in order to meaningfully address the moral chunking subtask, AU, in its pursuit of an SHMS integrating submitted candidates, must try to accommodate the potential breadth and variety of widespread human morality. Parameters must be specifiable (which means they do not need to be but can optionally be specified) not only for the agent (as e.g. in virtue ethics), not only for the action (as e.g. in deontology), not

https://www.wageningenacademic.com/doi/pdf/10.3920/978-90-8686-922-0_4 - Tuesday, October 04, 2022 6:24:32 AM - Utrecht University IP Address:131.211.12.11

only for the consequences (as e.g. in utilitarianism) but also for the first-person perspective of the perceiver performing the moral judgement (as e.g. in secular care ethics, in some schools of Indian philosophy but also in Christian ethics), which adds further relational considerations. But since the human perceiver exists within unique socio-psycho-techno-physical strata, has a unique developmental history, and is located in a specific spatial and temporal context, parameters as heterogeneous as socio-cultural background, psychological features, peculiarities in mind perception, current physiological and health state, and environmental conditions can have a decisive effect on the judgement of moral events (see also e.g. Haidt (2001) and Schwartz (2016) for more related perspectives on human ethical decision-making). Interestingly, even the language in which a moral event is presented can have an impact on moral judgements via differences in affective associations, which can lead to a tendency for more distance in second-language contexts (McHugh *et al.*, 2021). Beyond that, similar crashes with self-driving cars were considered more severe compared to those involving human drivers irrespective of cause (Liu *et al.*, 2019).

To sum up, it becomes clear that for sense-making regarding human morality across multiple different ethical frameworks, AU-based moral programming must at the minimum be able to accommodate a variety of parameters with regard to the following generic elements that can provisionally serve as an SHMS basis for moral chunking: perceiver, agent, action, patient. These four elements build up a tentative SHMS moral chunking tetrad for AU (hereafter referred to as SHMS chunking tetrad). That being said, it is important to note that, in addition, a breadth of epistemically relevant biases can arise in the context of moral judgements. This includes especially mind perception distortions (Gray *et al.*, 2012; Wegner and Gray, 2017) and formal inconsistencies. For instance, in empirical studies, people based moral judgements of artificial agents, among others on perceived intentionality and anthropomorphism (Bigman *et al.*, 2019), perceived feelings (Yam *et al.*, in press), and perceived agency, whereby the latter can even affect the sense of agency of the human perceiver (Ciardo *et al.*, 2020).

Instead of imposing a choice, AU provides a supportive generic interface as scaffold for candidate ethical frameworks. However, AU encourages critical thinking and flags epistemic issues related to (formal) inconsistencies that can affect AI safety post-deployment. For instance, for AI in ethical high-stake contexts, it has been shown that utilitarian and consequentialist utility functions face safety-relevant impossibility theorems (Eckersley, 2018), which can be derived formally when applying utilitarianism to population ethics (Arrhenius, 2000; Greaves, 2017). These impossibility theorems are linked to a mind perception problem that has been called 'the perspectival fallacy of utility

https://www.wageningenacademic.com/doi/pdf/10.3920/978-90-8686-922-0_4 - Tuesday, October 04, 2022 6:24:32 AM - Utrecht University IP Address:131.211.12.11

assignment'[3] (Aliman, 2020; Aliman and Kester, 2019a). Hence, AU flags classical utilitarian utility functions as problematic for high-stake AI. To solve this utility function problem in AI safety, different solutions have been proposed in recent years, among which the two options that we briefly enumerate next. As a first option, it is notable that AU itself facilitates the framing of utility functions at a higher abstraction level that is not subject to the utilitarian impossibility theorems. In fact, AU gains its name from this point of departure. Namely, in its quest for an SHMS, it was also considered whether ethical mathematically formulated functions such as AU candidates should always be problematic on theoretical grounds or whether a meaningful alternative was possible. Classical utilitarianism assigns utilities to states of the world, i.e. to time snapshots encoding merely the consequences of actions on the patient. Extending beyond that, AU facilitates the crafting of utility functions that are not affected by those impossibility theorems. This is possible under AU via dynamically updatable utility functions formulated at the level of time integrals covering the entire SHMS chunking tetrad, i.e. with information on perceiver, agent, action, and patient (Aliman, 2020). A second option to bypass utilitarian impossibility theorems is to probabilistically encode moral uncertainty into AI utility functions (which mathematically means to consider functions that are either partial orders or probability distributions over total orders instead of the classical total orders that utilitarianism uses) as advanced by Eckersley (2018). The latter leads us to the next subtask pertaining to the inescapable question: what if one is uncertain about one's ethical framework?

dynamically updatable at the level of time integrals

4.1.2 Epistemic dizziness

Under AU, it is not only acknowledged that one could be wrong in the context of moral decision-making but that one could always be wrong. This is in line with the epistemic assumptions of Karl Popper that all knowledge is fallible (Popper, 1995). Human beings experience embodied and socially modulated, cognitive-affective perspectival projections of the world. Also, under a deflationary account where truth is not equated with consensus, humans do not inhabit a 'post-truth' era (Aliman and Kester, 2020) since it is clear that even in the past sensory experience did not directly provide access to truth by virtue of being embedded in a web of prior knowledge from within an egocentric first-person perspective of the world shaped by cognitive-affective dynamics – making 'post-truth' a misnomer. Generally, well-tested unfalsified conjectures could be wrong, and previously empirically falsified conjectures could be true if the assumptions underlying the empirical observations were wrong (Frederick, 2020a). Affective realism, the phenomenon that one tends to

[3] For exemplary scenarios on how this fallacy manifests, see Aliman, 2020.

https://www.wageningenacademic.com/doi/pdf/10.3920/978-90-8686-922-0_4 - Tuesday, October 04, 2022 6:24:32 AM - Utrecht University IP Address:131.211.12.11

see what one believes via affect being interwoven in perception (Fridman *et al.*, 2019; Siegel *et al.*, 2018), adds up to what we call epistemic dizziness. Morality is of a conjectural nature; even unconscious components stem from a generative model of the world with certain priors encoding assumptions and associations from the past, which is key to affective realism. In short, morality too is affected by this inescapable epistemic dizziness. Fortunately, our general epistemic aim can still be to create ever better explanations (Frederick, 2020a), i.e. to tease out better from worse explanations (on the basis of theoretical criticism and if possible empirical falsification) while steadily creating new ones. The crucial question now becomes: what does it signify for moral programming?

Firstly, due to the fallibility of human knowledge, any framework for meaningful AI control must be updatable-by-design. (Since one cannot scientifically predict the future of knowledge creation, even entities that claim to know objective moral truth today cannot predict that they would not be able to change their own view in the future.) When applied to an AI that is deployed on societally relevant scales, such a generic update mechanism has been termed to instantiate a socio-technological feedback loop (Aliman, 2020). Secondly, the starting point for such an AI-enacted moral feedback mechanism (even if on smaller scales) need not and cannot logically be proven right. Justifications are logically invalid, as shown by Popper long ago (Bartley, 1976). For instance, when contemplating the statement that all swans are white, no amount of repeated observations of white swans can prove its veracity. Hence, it is sufficient to choose reasonable, well-tried starting points without truth-related claims. Depending on the scale of the AI deployment, one can select from a vast array of internally formally consistent options comprising e.g. normative theories from one's culture, the AU-enhanced tetradic utility function scaffold (SHMS) filled in with parameters from a self-selected ethical framework, legal rules for criminalised contexts, some parameters from the UN Sustainable Developmental Goals, other international norms, comparative theories, pragmatic alternatives, any tailored combinations of those or weighted sums, and so forth. Thirdly, AU can then be utilised to refine moral reasoning if desired. For instance, with the AU-enhanced utility function option, one can test AI-enacted moral knowledge specified at the beginning of an iteration of a socio-technological feedback loop or organisational feedback loop. Using simulation environments, one can exploit counterfactual reasoning (Aliman and Kester, 2019a) and craft adversarial AU-enhanced utility functions either to let different moral conceptions compete against each other for self-education or purely in order to explore novel candidates for one's own moral values.

The question at the end of the last section asked: what if one is uncertain about one's ethical framework? Two remaining caveats relevant to answering this question are addressed in this and the next paragraph. Firstly, irrespective of

https://www.wageningenacademic.com/doi/pdf/10.3920/978-90-8686-922-0_4 - Tuesday, October 04, 2022 6:24:32 AM - Utrecht University IP Address:131.211.12.11

the ethical framework chosen, there is in general an inherent limitation to the injection of moral models in AI systems. In fact, depending on the framework, one assumes pre-deployment that the entirety of options, alternatives, parameters, and/or outcomes is known in advance. However, human moral decision-making is unpredictable, often disruptive, and can even be adversarial as one can deduce from the example with the adversarial AU-enhanced utility function. As stated by Deutsch (2011): 'It is a mistake to conceive of choice and decision-making as a process of selecting from existing options according to a fixed formula. That omits the most important element of decision-making, namely the creation of new options.' In short, in moral decision-making, it can be that people create additional yet unknown alternatives. It can be that they delete and replace initial moral assumptions if they adversarially experiment with novel differing values that they create or discover. From this point of view, 'rationality permits us to act in accord with our best-tested theories, since they may be true; but it also permits us to act against them, precisely because our best-tested theories may be false and may, indeed, be refuted when we act against them' (Frederick, 2020b). A deployed present-day AI is not able to perform a creative adversarial task requiring explanatory knowledge. It is for this reason that we depicted AU-based moral programming as facilitating the design of heuristic moral models, ephemeral approximative shadows of morality to be necessarily updated with time.

heuristic moral models

Secondly, the idea of crafting moral models by assigning probabilistic uncertainty to moral values themselves (which is different from weighted sums) in the hope of outsourcing its resolution to an AI system post-deployment brings about epistemic and safety-relevant caveats. From epistemic dizziness it is clear that objective probabilities for conjectures, i.e. absolute certainty, are never guaranteed for lack of experiencing truth directly. However, the case of subjective probabilities for normative ethical frameworks can lead to self-contradictory implications for AI, among others because it is epistemically permissible to act against one's prior assumptions (Frederick, 2020a) (e.g. in an attempt to test them). Therefore, this also includes: (1) the option to act against one's own subjective probabilities; and (2) to act on novel differing spontaneously created options. The former can lead to AI control *ad absurdum* and the latter cannot be performed by present-day AI since explanatory knowledge is required. In short, probabilistic moral uncertainty in moral programming is feasible if seen as heuristic, and whose resolution would require the AI to prompt a human at the post-deployment stage. Otherwise, especially in ethical high-stake cases, to delegate a metaphorical throwing of (crooked) dice (or flipism) to a deployed AI system only adds blurriness to the already doubt-ridden heuristic moral models, is not optimal for cybernetic control (simply put, incomplete knowledge is better than none (Wiener, 1960)), and can be highly questionable for reasons of safety, accountability, and lack of

https://www.wageningenacademic.com/doi/pdf/10.3920/978-90-8686-922-0_4 - Tuesday, October 04, 2022 6:24:32 AM - Utrecht University IP Address:131.211.12.11

participatory decision-making. The existence of epistemic dizziness means we cannot inherently eliminate doubt from the world, but we need not magnify it. We can accept the conjecture that we could always be wrong but can strive for ever better (moral) explanations within (moral) feedback-loop mechanisms, among others to craft better AI-enacted heuristic moral models. In a nutshell, uncertain humans equipped with some dice at the time of moral decision-making could throw those dice but could also unexpectedly (co-)create novel as yet unknown solutions for the problem – something present-day AIs cannot. Hence, in ethical high-stake cases of moral programming, it is recommended to reserve the throwing of dice (if really necessary) to people.

4.2 Practical use of theoretical solution

As mentioned in Section 4.1, AU-based meaningful AI control can be framed in terms of three subtasks. In this section, we will finally address the third one. After integrating theoretical knowledge on how to tackle the first two subtasks, i.e. the moral chunking problem and epistemic dizziness, one is still faced with the complex systems engineering problem of devising safety-aware strategies for AU-based AI deployment. Against the backdrop of AI risks in the form of first-order harm and potential second-order repercussions emerging amidst interwoven multi-stakeholder settings – be it at the level of an organisation or larger societal contexts – this third AU subtask meets non-trivial challenges. In this section, we elucidate how the theoretical foundations from the last section can be harnessed for moral programming management in practical settings with AU as an interface. Finally, we briefly touch upon a practical open question whose relevance has increased in recent years: can we craft a principled approach against the risks posed by deploying AI operating in regimes that widely exceed their human counterparts with regard to speed, scale, and selectivity? We deconstruct the underlying issue, introduce what we call the 'no-body problem' in AI safety and explain how one possible option could consist of utilising AU in conjunction with a cybernetically motivated strategy for advanced AI safety; however, we also depict novel caveats.

systems engineering problem

4.2.1 Moral programming management with augmented utilitarianism

For small-scale AI deployment contexts, AU can already fulfil a relevant function in the design process. Namely, in small-scale AU-based AI design the domain-general requirement is twofold. Firstly, for safety reasons, there is a vital requirement for only offering well-formed internally consistent machine-readable heuristic moral models as options, i.e. that are for instance not known to be affected by mathematically specified impossibility theorems. Conceivable machine-readable formats range from rule-based forms to AU-enhanced utility

https://www.wageningenacademic.com/doi/pdf/10.3920/978-90-8686-922-0_4 - Tuesday, October 04, 2022 6:24:32 AM - Utrecht University IP Address:131.211.12.11

functions and to knowledge graphs. Secondly, there is the requirement for designers to provide an updatable and user-adjustable interface giving each user the possibility (if desired) to tailor their own individualised heuristic moral model selecting from a wide pool of specific elements sampled from the entire generic SHMS chunking tetrad. Importantly, since AU is non-normative, meta-ethical, and descriptive, it only acts as an empty structured scaffold left blank. It is the user who then specifies whether and which of the multiple thinkable and available moral parameters related to perceiver, agent, action, and/or patient to instantiate. As a result, via the tailored AU encoding, the problem-solving ability of the AI system can now operate within the self-determined individualised moral bounds of the user. Optionally, designers can in addition provide a pre-selection, e.g. from common heuristic moral models or their own inclinations.

For AI deployed in public domains and in large-scale legally relevant contexts, the emerging accountability issues require a clean assignment of responsibilities. In this case, the well-formed machine-readable heuristic moral models should ideally, and where feasible, be constrained not only by soft law but also by hard law. However, it seems worth acknowledging that a central AI safety regulation could – if not actively counteracted with novel practical measures – undesirably entail bureaucratic layers that are temporally not on a par with ongoing technological progress. In any case, an inherent requirement of a scheme compatible with either hard or soft law (as was also the case in the small-scale setting) is that the fine-tuning of a heuristic moral model and the design of an AI system focusing on its problem-solving ability in a given domain are separable. More precisely, for such a strategy to be feasible one requires AI architectures for which goals and problem-solving ability are orthogonal to each other (i.e. can be freely combined). Only via this type of orthogonality-based disentanglement can the described AU-based strategy be realised. (Note that orthogonality-based disentanglement is decisively different from and not to be confused with the orthogonality thesis propounded by Bostrom claiming that 'intelligence and final goals are orthogonal axes along which possible agents can freely vary' (Bostrom, 2012). However, by contrast it holds that: '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' (Aliman, 2020).) From a system engineering perspective this signifies that there must be an AI system architecture for which one is able to separate the 'what' from the 'how' in order to allow for a disentangled assignment of responsibilities. Examples of existing types of AI systems where this is possible are deep-learning systems and hybrid intelligent systems instantiating MDPs

https://www.wageningenacademic.com/doi/pdf/10.3920/978-90-8686-922-0_4 - Tuesday, October 04, 2022 6:24:32 AM - Utrecht University IP Address:131.211.12.11

in utility-based planning (but also the more recent active inference agents (Smith *et al.*, 2021)). Intelligent systems are able to perform the OODA loop and have been often referred to as 'autonomous systems´. However, we choose to use the term 'intelligent system' since those AIs are not setting their own goals autonomously. Instead, they always act on human-defined heuristic moral models.

In Figure 4.1, we provide a simplified illustration of deploying an intelligent system in large-scale contexts utilising a stakeholder-crafted AU-encoding constrained – in this example – by hard law. The stakeholder in question is a societally relevant entity performing a certain operation. For transparency and to improve explainability, the intelligent system provides information related to its internal working via counterfactuals of selected courses of action to clarify its enactment towards surroundings and stakeholder. If later required, these samples of counterfactual reasoning can be used for auditing and in forensic and judicial contexts. The manufacturers are responsible for conducting safety and security tests pre-deployment and throughout the lifecycle. Importantly, for reasons of formal inconsistencies discussed earlier, if the chosen AU-encoding represents a utility function, it cannot be a conventional one. It must be an alternative not prone to impossibility theorems, such as e.g. AU-enhanced utility functions (where norms and rules can also be used as constraints on the function), an inter-theoretic weighted sum option or morally uncertain functions with partial orders necessarily leading to prompt human entities. On the whole, the figure contains multiple feedback loops. While an in-depth analysis is beyond the scope of this chapter and has been performed elsewhere, we draw the attention of the reader to two AU-based feedback loops. The outer feedback loop highlighted in blue comprising legislative power, stakeholder, intelligent system, and environment reflects a socio-technological feedback loop. At that level, which can also be supported by simulations or small experiments in synthetic environments, a societal level ethical enhancement can take place by testing and refining heuristic moral models in practice as discussed. However, the small inner feedback-loop between stakeholder and intelligent system (and analogically between user and AI in small-scale settings), while apparently simple, already reflects an emerging complex issue for the deployment of future more advanced AI surpassing human entities with regard to speed, scale, and scope. While in smaller less complex contexts harm incurred by AI seems more manageable, it is thinkable that future AI deployment could yield more unforeseen repercussions. In the following subsection, we briefly elucidate how a further analysis of this feedback loop reveals what we denote as the 'no-body problem'.

https://www.wageningenacademic.com/doi/pdf/10.3920/978-90-8686-922-0_4 - Tuesday, October 04, 2022 6:24:32 AM - Utrecht University IP Address:131.211.12.11

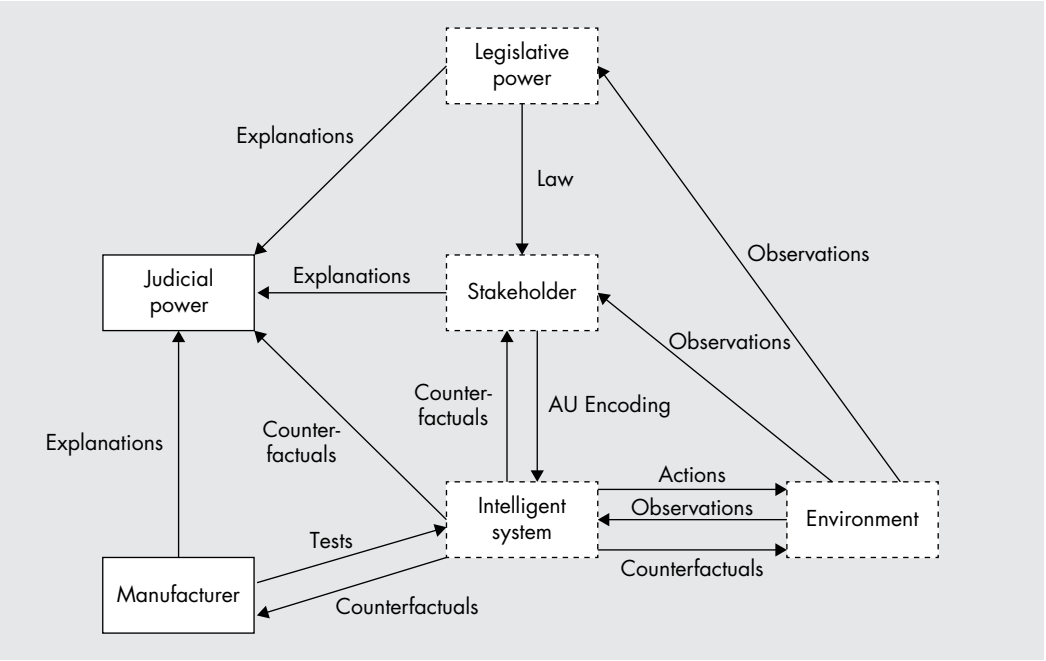

Figure 4.1. Simplified illustration of an AU-based socio-technological feedback loop (highlighted with dashed lines).

4.2.2 Advanced Type I artificial intelligence risks and the 'no-body problem'

Before discussing AI risks, we taxonomically distinguish between two types of systems. Unlike humans, present-day AI cannot consciously create and understand explanatory knowledge. Recently, the substrate-independent term 'Type II system' has been proposed (Aliman, 2020) for systems that are able to consciously create and understand explanatory knowledge. Type I systems are all systems that are not able to fulfil this task. Hence, from that perspective, all present-day AIs are Type I and non-conscious, and there is no single existing Type II AI. Non-human animals are conscious Type I systems. More broadly, humans are currently the only known instances of Type II systems (all of which are necessarily conscious). However, from a purely theoretical standpoint, while no such artificial system has been implemented anywhere on earth, Type II AI must be possible since it is not prohibited by any law of nature and due to the universality of computation. The AI safety paradox states that control and value alignment are conjugate requirements in AI safety. To put it plainly, you cannot simultaneously control and value-align with the same entity. Value alignment via morality necessitates understanding and explanatory knowledge which Type I AI is not able to. In other words, although the current aim of many researchers is to achieve the conjunction of control and value alignment for

https://www.wageningenacademic.com/doi/pdf/10.3920/978-90-8686-922-0_4 - Tuesday, October 04, 2022 6:24:32 AM - Utrecht University IP Address:131.211.12.11

Type I AI, it is postulated that we can try to control but cannot genuinely value-align with it across our domains of interest. Once a Type II AI exists, we could then value-align with it (if it agrees), but we could not control it, in the same way that we cannot control humans in the long term and slavery is unethical. This holds for humans as well as for hypothetical Type II AI. Importantly, it does not make sense to say that a Type II system is more dangerous than a non-conscious Type I system. It would be analogous to comparing humans with knives. Humans and also hypothetical Type II AI could of course have violent intentions, but it is not what determines their Type II nature. Malicious actors could already now design hazardous Type I AI to provoke serious risks. (In Chapter 10, we discuss an example of immoral programming with current language AI and the significant related risks.)

Generally, the Type II AI case is not of practical relevance nowadays and its issue would be a hybrid case of participatory sense-making amongst explanatory knowledge creators. Hence, in the following, we focus strictly on existing and thus non-conscious Type I AIs whose abilities surpass humans in scale, speed, and/or selectivity. Arguably, intelligent systems which were depicted in the small feedback loop with stakeholders/users are candidates for new emerging advanced Type I AI risks. An interesting question then becomes: how can advanced Type I AI control be implemented without real value alignment? In our view, zooming into this human vs Type I AI feedback loop may reveal some clues about the underlying cybernetically relevant message transmission problem. In 1960, Norbert Wiener, one of the first cyberneticians, stated that: 'Human action is a feedback action' (Wiener, 1960). In fact, it is insufficient that humans would in theory be able to react appropriately to a Type I AI's message in the specified form of counterfactuals. If it takes humans too long to understand that message in the first place, unforeseen undesirable consequences might already have occurred before they were able to act on it, e.g. by manually turning it off. Wiener gave the hypothetical example of an automated bottle factory programmed to maximise on productivity and whose owner could be 'made bankrupt by the enormous inventory of unsalable bottles manufactured before he learns he should have stopped production six months earlier' (Wiener, 1960). It is on such grounds that many AI safety researchers conjecture about the need to make Type I AI artificially stupid (Trazzi and Yampolskiy, 2020), i.e. to artificially maintain AI abilities at human level or below before it reaches 'superintelligence' and outsmarts human counterparts. But in our view, 'the price of security is eternal creativity' (Aliman *et al.*, 2021). From a functional viewpoint, considering the different ways of information processing, one could consider humans and Type I AI as disparate subparts of the same functional unit where the issue is to add meaning to the messages and synchronise their transmission in order to maintain explanatory knowledge creation as the most relevant subset of creativity. The substrate-independent

https://www.wageningenacademic.com/doi/pdf/10.3920/978-90-8686-922-0_4 - Tuesday, October 04, 2022 6:24:32 AM - Utrecht University IP Address:131.211.12.11

construct of a hybrid functional unit comprising Type II entities (today only applicable to humans) and Type I entities (such as Type I AIs but also any other Type I entities) can be termed a cyborgnet. (The latter is essentially different from the word cyborg, is formulated at a much more general level and embedded in an own novel theoretical framework (Aliman, 2021).) We explain how this could finally motivate one possible future strategy.

cyborgnet

In a cyborgnet, creativity is necessarily embodied since Type II entities are also physical entities. From this point of view, a communication between humans and Type I AI that does not adjust to human bodily elements does not contain enough information. To put it very simply, most present-day heuristic moral models for Type I AI belong to no-body. Applying our functional stance, it would be embodied cyborgnetic creativity that one intends to augment and not the intelligence of an AI separately or the intelligence of an isolated human separately. In short, the body of relevance to consider is the 'body' of the entire cyborgnet in a given deployment context. A future Type I intelligent system that would itself be deployed in robotic settings may be equipped with technically self-controlled (cyber-)physical elements. However, it would then not suffice for it to develop a technical self-assessment and self-management of those own parts, but one would also integrate a model of the human physical entity in real time. This could include Type-I-AI-directed voice commands, a hybrid mix of context-sensitive experience sampling via physiological activity including arousal monitoring, human valence ratings, and so forth. Feedback loops are by definition bi-directional. Humans could receive additional information on the state of the Type I AI. For instance, next to the counterfactuals projected to past and future, information of the physical state of the Type AI could be visualised efficiently or conveyed via sensory extensions. This could as a side-effect create novel socio-material affordances stimulating human creativity. To sum up, solving the no-body problem bi-directionally via conceiving of humans and Type I AI as disparate but connected entities within one cyborgnet engaging in feedback loops could improve the underlying intra-cyborgnet message transmission and provide a physical grounding of meaning. In turn, this might mitigate the Type I AI control issue by transforming it into a bi-directional hybrid functional integration task at the level of a cyborgnet. Ultimately, one could for instance inject real-time human bodily signals as an additional parameter into an AU-based heuristic moral model. In this way, the problem-solving ability of Type I AI could be funnelled not only by human morality but also by embodied cyborgnetic creativity. (In Chapter 10, we discuss how present-day AI could be employed to counteract AI attacks themselves via cyborgnetic creativity augmentation.) Hence, instead of designing an isolated Type I AI made artificially stupid with human intelligence as upper-bound, this cyborgnetic functional design would strive to ideally attune and tailor its subcomponents to each other – the body included. And as is often said, the

https://www.wageningenacademic.com/doi/pdf/10.3920/978-90-8686-922-0_4 - Tuesday, October 04, 2022 6:24:32 AM - Utrecht University IP Address:131.211.12.11

whole is more than the sum of its parts... On the downside, a higher capacity information channel (given that more data is transferred) may yield higher costs. Moreover, privacy concerns and security risks may arise. Private data shared in those cyborgnetic information exchange networks could be exploited by malevolent actors. However, to counteract these risks, the zero-trust paradigm from cybersecurity must be proactively applied at multiple levels.

4.3 Conclusions

Wernaart (2021) delineated a sophisticated starting point for a roadmap on moral programming including ethical high-stake cases where the designer is not the central moral authority. AU is one possible non-normative meta-ethical explanatory framework that can be utilised for moral programming in a pluralistic society. In short, the aim of the transdisciplinary AU framework is to accommodate for moral pluralism requirements by providing a supportive generic SHMS scaffold left blank, in which moral authorities (especially users or society but also designers in default settings) fill in flexible updatable and machine-readable heuristic moral models. We explained how AU tackles the three main subtasks of this complex endeavour: the moral chunking problem, epistemic dizziness, and the practical systems engineering subtask of AU-based AI deployment. Finally, we offered a differentiated account of advanced AI risks and reflected on a cyborgnetic embodiment strategy that could yield resilience – albeit not immunity – providing a variety of incentives for future work.

References

Aliman, N.M., 2020. Hybrid cognitive-affective strategies for AI safety. Doctoral dissertation, Utrecht University, the Netherlands. https://doi.org/10.33540/203

Aliman, N.M., 2021. Cyborgnet theory. Available at: https://nadishamarie.jimdo.com/app/download/10822805671/CT.pdf?t=1632837017.

Aliman, N.M. and Kester, L., 2019a. Transformative Al governance and Al-empowered ethical enhancement through preemptive simulations. Delphi – Interdisciplinary Review of Emerging Technologies, 2: 23-29. https://doi.org/10.21552/delphi/2019/1/6

Aliman, N.M. and Kester, L., 2019b. Requisite variety in ethical utility functions for AI value alignment. In Workshop on Artificial Intelligence Safety 2019 co-located with the 28th International Joint Conference on Artificial Intelligence AISafety@IJCAI 2019. Available at: http://ceur-ws.org/Vol-2419/paper_12.pdf.

Aliman, N.M. and Kester, L., 2020. Facing immersive 'post-truth' in AIVR?. Philosophies, 5: 45. https://doi.org/10.3390/philosophies5040045

Aliman, N.M., Kester, L. and Yampolskiy, R., 2021. Transdisciplinary AI observatory – retrospective analyses and future-oriented contradistinctions. Philosophies, 6: 6. https://doi.org/10.3390/philosophies6010006

https://www.wageningenacademic.com/doi/pdf/10.3920/978-90-8686-922-0_4 - Tuesday, October 04, 2022 6:24:32 AM - Utrecht University IP Address:131.211.12.11

Arrhenius, G., 2000. An impossibility theorem for welfarist axiologies. Economics & Philosophy, 16: 247-266.

Barrett, L.F., 2017. How emotions are made: the secret life of the brain. Houghton Mifflin Harcourt, Boston, MA, USA.

Bartley, W.W., 1976 The philosophy of Karl Popper. Philosophia 6: 463-494. https://doi.org/10.1007/BF02379286

Bigman, Y.E., Waytz, A., Alterovitz, R. and Gray, K., 2019. Holding robots responsible: the elements of machine morality. Trends in cognitive sciences, 23: 365-368. https://doi.org/10.1016/j.tics.2019.02.008

Bostrom, N., 2012. The superintelligent will: motivation and instrumental rationality in advanced artificial agents. Minds and Machines, 22: 71-85. https://doi.org/10.1007/s11023-012-9281-3

Ciardo, F., Beyer, F., De Tommaso, D. and Wykowska, A., 2020. Attribution of intentional agency towards robots reduces one's own sense of agency. Cognition, 194: 104109. https://doi.org/10.1016/j.cognition.2019.104109

Deutsch, D., 2011. The beginning of infinity: explanations that transform the world. Penguin, London, UK.

Dubljević, V., List, G.F., Milojevich, J., Ajmeri, N., Bauer, W., Singh, M.P., Bardaka, E., Birkland, T., Edwards, C., Mayer, R. and Muntean, I., 2021. Toward a rational and ethical sociotechnical system of autonomous vehicles: a novel application of multi-criteria decision analysis. arXiv preprint arXiv:2102.02928 https://arxiv.org/abs/2102.02928

Dubljević, V., Sattler, S. and Racine, E., 2018. Deciphering moral intuition: how agents, deeds, and consequences influence moral judgment. PloS one, 13: p.e0204631. https://doi.org/10.1371/journal.pone.0206750

Eckersley, P., 2018. Impossibility and uncertainty theorems in AI value alignment (or why your AGI should not have a utility function). arXiv preprint arXiv:1901.00064. https://arxiv.org/abs/1901.00064

Frederick, D., 2020a. Against the philosophical tide: essays in Popperian critical rationalism. Critias Publishing. Yeovil, UK.

Frederick, D., 2020b. Falsificationism and the pragmatic problem of induction. Organon F, 27: 494-503.

Fridman, J., Barrett, L.F., Wormwood, J.B. and Quigley, K.S., 2019. Applying the theory of constructed emotion to police decision making. Frontiers in Psychology, 10: 1946. https://doi.org/10.3389/fpsyg.2019.01946

Gray, K., Young, L. and Waytz, A., 2012. Mind perception is the essence of morality. Psychological Inquiry, 23: 101-124. https://doi.org/10.1080/1047840X.2012.651387

Greaves, H., 2017. Population axiology. Philosophy Compass, 12: p.e12442. https://doi.org/10.1111/phc3.12442

Haidt, J., 2001. The emotional dog and its rational tail: a social intuitionist approach to moral judgement. Psychological Review, 108(4), 814-834. https://doi.org/10.1037/0033-295X.108.4.814

https://www.wageningenacademic.com/doi/pdf/10.3920/978-90-8686-922-0_4 - Tuesday, October 04, 2022 6:24:32 AM - Utrecht University IP Address:131.211.12.11

Hoemann, K. and Feldman Barrett, L., 2019. Concepts dissolve artificial boundaries in the study of emotion and cognition, uniting body, brain, and mind. Cognition and Emotion, 33: 67-76. https://doi.org/10.1080/02699931.2018.1535428

Isbilen, E.S. and Christiansen, M.H., 2020. Chunk-based memory constraints on the cultural evolution of language. Topics in Cognitive Science, 12: 713-726. https://doi.org/10.1111/tops.12376

Kleckner, I.R., Zhang, J., Touroutoglou, A., Chanes, L., Xia, C., Simmons, W.K., Quigley, K.S., Dickerson, B.C. and Barrett, L.F., 2017. Evidence for a large-scale brain system supporting allostasis and interoception in humans. Nature Human Behaviour, 1: 1-14. https://doi.org/10.1038/s41562-017-0069

Liu, P., Du, Y. and Xu, Z., 2019. Machines versus humans: people's biased responses to traffic accidents involving self-driving vehicles. Accident Analysis & Prevention, 125: 232-240. https://doi.org/10.1016/j.aap.2019.02.012

McHugh, C., McGann, M., Igou, E.R., and Kinsella, E.L., 2021. Moral judgment as categorization (MJAC). Perspectives on Psychological Science. https://doi.org/10.1177/1745691621990636

Oosterwijk, S., Lindquist, K.A., Anderson, E., Dautoff, R., Moriguchi, Y. and Barrett, L.F., 2012. States of mind: emotions, body feelings, and thoughts share distributed neural networks. NeuroImage, 62: 2110-2128. https://doi.org/10.1016/j.neuroimage.2012.05.079

Popper, K.R. 1995. In search of a better world: Lectures and essays from thirty years. Routledge, Brighton, UK.

Schein, C. and Gray, K., 2018. The theory of dyadic morality: reinventing moral judgment by redefining harm. Personality and Social Psychology Review, 22: 32-70. https://doi.org/10.1177/1088868317698288

Schwartz, M.S., 2016. Ethical decision-making theory: an integrated approach. Journal of Business Ethics, 139: 755-776. https://doi.org/10.1007/s10551-015-2886-8

Siegel, E.H., Wormwood, J.B., Quigley, K.S. and Barrett, L.F., 2018. Seeing what you feel: affect drives visual perception of structurally neutral faces. Psychological Science, 29: 496-503. https://doi.org/10.1177/0956797617741718

Smith, R., Friston, K. and Whyte, C., 2021. A step-by-step tutorial on active inference and its application to empirical data. https://psyarxiv.com/b4jm6/

Trazzi, M. and Yampolskiy, R.V., 2020. Artificial stupidity: data we need to make machines our equals. Patterns, 1: 100021. https://doi.org/10.1016/j.patter.2020.100021

Wegner, D.M. and Gray, K., 2017. The mind club: who thinks, what feels, and why it matters. Penguin, London, UK.

Wernaart, B., 2021. Developing a roadmap for the moral programming of smart technology. Technology in Society, 64: 101466. https://doi.org/10.1016/j.techsoc.2020.101466

Wiener, N., 1960. Some moral and technical consequences of automation. Science, 131: 1355-1358. https://www.jstor.org/stable/1705998

Yam, K.C., Bigman, Y.E., Tang, P.M., Ilies, R., De Cremer, D., Soh, H. and Gray, K., in press. Robots at work: people prefer – and forgive – service robots with perceived feelings. The Journal of Applied Psychology, ###. https://doi.org/10.1037/apl0000834