Values in science and AI alignment research

Leonard Dung (leonard.dung@rub.de)

Abstract

Roughly, empirical AI alignment research (AIA) is an area of AI research which investigates empirically how to design AI systems in line with human goals. This paper examines the role of non-epistemic values in AIA. It argues that: (1) Sciences differ in the degree to which values influence them. (2) AIA is strongly value-laden. (3) This influence of values is managed inappropriately and thus threatens AIA's epistemic integrity and ethical beneficence. (4) AIA should strive to achieve value transparency, critical scrutiny from inside and outside the discipline – involving the public –, and to empower actors without strong commercial interests.

1. Introduction

In a nutshell, artificial intelligence alignment research (henceforth "AIA") is an area of AI research which investigates how to design AI systems in line with human values. In this paper, I will focus on AIA research which consists in empirical studies on actual AI systems or is tightly related to such empirical work. I will set aside alignment research which is based on formal mathematical or informal conceptual reasoning, including research on ethical questions. While empirical AIA is a very recent endeavor, there already exists a large number of publications, which – moreover – is rapidly growing (for a comprehensive review, see Ji et al. 2023).

In this paper, I analyze the role of non-epistemic values in AIA. It is clear that epistemic values, i.e. values indicative of truth and knowledge (such as empirical adequacy or simplicity), figure in AIA. However, the presence and extent of non-epistemic values in AIA is an open question. In section 2, I begin by providing a cursory and selective overview of AIA as a scientific field. This will involve reviewing discussions of what "alignment" means and concrete empirical discoveries from recent research. Then, this paper makes four substantial contributions: First (section 3), I argue that sciences differ in the degree to which values influence them. Second, I show that AIA is especially – strongly – value-laden. Third (section 4), I identify risks to the epistemic and ethical integrity of AIA which stem from its value-ladenness. Fourth, I develop proposals for how values in AIA can be managed properly. By making these contributions, I suggest that – despite common claims that AIA is a "pre-paradigmatic" field (Kirchner et al. 2022; Kuhn 1962; Ze Shen 2022) which lacks established

and agreed-upon theories, methods, and concepts – at least some strands of empirical AIA are sufficiently unified and distinctive to allow for fruitful investigation by philosophers of science. Section 5 concludes.

2. AI alignment research

2.1 Alignment: multiple concepts

Since its inception, AIA is motivated by concern about the societal impacts of AI (Yudkowsky 2016). Such impacts encompass both current harms as well as anticipated risks, including more speculative catastrophic and existential risks (Bostrom 2014; Dung 2024a). AI alignment is frequently seen as the solution for, or at least helpful to, these concerns (Christian 2020; Russell 2019; Shevlane et al. 2023). So, what does "AI alignment" mean? I will not propose a full conceptual analysis, since I am ultimately concerned with the empirical research conducted under the label "alignment", not with legislating the meaning of the term. However, some conceptual distinctions are useful. First, let us distinguish *ethical* and *technical* alignment (Gabriel 2020). Technical alignment requires building AI systems such that they try to what their designers want them to do, i.e., building them such that they robustly pursue the intended goals (Dung 2023). So, the system needs to be interpretable as goal-directed, even though it may lack goals in a substantive sense (Butlin 2023; Dung 2024b), and these goals need to conform to what their designers intend. Ethical alignment requires selecting the ethically correct goals to align AI systems with. I will focus in this paper on alignment research conducted with scientific methods, since it is obvious that ethical alignment research directly involves ethical values. Scientific alignment research examines technical alignment (some qualifications to this claim follow later), so henceforth "alignment" will mean "technical alignment", unless otherwise noted.

Alignment is often contrasted with *capability*. Intuitively, if a system does not do what it is intended to do (e.g., win a video game), this may be because it is misaligned (it does not try to win the game, but pursues some other goal (e.g. OpenAI 2016)) or because it lacks capability (it is not skilled enough to win the game) (Dung 2023). Very capable but misaligned systems are typically taken to be especially dangerous. Also, researchers often distinguish "outer" from "inner" alignment (e.g., Hubinger 2020), where outer alignment requires an appropriate specification of the intended goal and inner alignment requires that the system actually, robustly, pursues this goal.[1] Finally, one can distinguish "prosaic" alignment, which

[1] In reinforcement learning, outer alignment can be contrasted with reward misspecification (Pan et al. 2022) and inner alignment with goal misgeneralization (Langosco et al. 2023).

concerns aligning current systems, with "superalignment", which concerns aligning hypothetical future systems whose capabilities exceed humans in most or all domains (Burns et al. 2023).

All these conceptual distinctions are potentially problematic. First, the distinction between alignment and capability seems blurry: current alignment techniques like reinforcement learning from human feedback (RLHF) (see below) also make the system more capable at tests of cognitive performance (Bai, Jones, et al. 2022; Bubeck et al. 2023). Second, it has been questioned whether the challenge of aligning AI can be fruitfully decomposed into inner and outer alignment (Hubinger 2021; Turner 2022). Third, it is debated to what extent techniques for prosaic alignment may generalize to super alignment, or whether the challenges of super alignment are too distinct and discontinuous (Casper et al. 2023; Cotra 2021; Dung 2023). The precariousness of the conceptual foundations of AIA presents an excellent opportunity for future philosophical inquiry. As of now, it is unclear whether AIA has a unified and distinctive subject matter about which it can produce generalizable causal claims. Partly, this is because it is unclear where exactly the borders of AIA are. However, I will set aside these issues here to single out specifically the role of non-epistemic values in AIA. In discussing AIA, I will focus on technical AIA and, more specifically, on studies similar to the paradigmatic examples cited below, without decisively settling which other cases of AI research count as AIA research.

2.2 Examples of current AIA

The goal of AIA is developing techniques for ensuring (technical) alignment of AI systems and producing generalizable knowledge about properties of (mis)alignment. Let us look at some examples. Since I will simplify to preserve space, I encourage the reader to consult the references for details.

In practice, RLHF is currently the dominant technique for aligning LLMs. In RLHF, human raters first rank different text completions of pre-trained LLMs according to some criteria (typically, how "helpful, harmless, and honest" they are (Bai, Jones, et al. 2022)). The resulting dataset is then used to train a reward model which numerically scores LLM outputs based on how well they meet these criteria. Finally, the LLM is trained via reinforcement learning (RL) to maximize the expected rewards from the reward model. Consequently, the weights of the LLM gradually adjust such that outputs which are assessed positively by human raters become more likely.

A prominent research direction concerns automating the process of providing feedback further. In constitutional AI (Bai, Kadavath, et al. 2022), LLMs are asked to critique their own responses to prompts based on a human-curated list of principles (a "constitution"), and to revise their original responses in light of this critique. After several iterations, a pretrained LLM is finetuned via supervised learning on the revised responses. Then, mimicking RLHF but with AI feedback, a further model is asked which out of several responses is best according to the constitution, and the resulting dataset is used to train a reward model which can then be used to train the LLM via RL.[2] Bai, Kadavath, et al. (2022) conclude that this method makes "it possible to control AI behavior more precisely and with far fewer human labels". In a different variation of standard RLHF, Korbak et al. (2023) find that LLM responses, as rated by humans, are more satisfactory – without impairment of task performance – when pre-training involves human feedback, rather than pre-training exclusively on text data and fine-tuning on human feedback afterwards.

Continuing the theme of automating the feedback process, Burns et al. (2023) show that, under some conditions, a more capable model (such as GPT-4) can be successfully finetuned on feedback by a less capable model (such as GPT-2), exceeding the performance of the weaker model. Burns et al. are interested in this phenomenon, which they call "weak-to-strong generalization", because it suggests that it may be possible to align future systems with superhuman capabilities based on feedback by less capable models (or humans).

In experiments on adversarial robustness (Schlarmann and Hein 2023; Zou, Wang, et al. 2023), researchers test whether it is possible to elicit undesirable inputs from systems (e.g., plans to build a bomb) with inputs specifically selected for this purpose, and try to make systems immune against such adversarial attacks. As a final example, Hubinger et al. (2024) train language models to produce undesirable behavior if and only if the prompt contains a certain specific trigger, creating a "model organism of misalignment". They show that standard alignment techniques, namely RLHF, finetuning on helpful, harmless, and honest outputs and adversarial training, are sometimes jointly insufficient to remove this "backdoor", which allows triggering undesirable behavior.

The preceding is an incomplete selection of important alignment techniques and results.[3] However, it is sufficient to show that AIA displays several characteristics of a scientific field: researchers investigate a distinctive set of alignment techniques, they empirically examine the

[2] Bai, Kadavath, et al. (2022) use a combination of human and AI feedback.

[3] See, e.g., representation engineering (Zou, Phan, et al. 2023) and other interpretability research (Pacchiardi et al. 2023) for further examples.

features of these techniques and make generalizable causal claims about them (e.g., Hubinger et al. (2024) show that standard RLHF, supervised finetuning, and adversarial training are sometimes insufficient to remove "backdoors" for undesirable behavior), and they build upon each other's work, making cumulative progress possible. In the next section, I will take this selective overview as a basis to elucidate the role of values in AIA.

3. Value-ladenness and AI alignment research

3.1 Uncontroversial forms of value influence

For our purposes here, we can say that a value is the belief that something is desirable or should be the case. My claim is that values play an especially pervasive and important role in AIA, compared to many other sciences. To give substance to this claim, different forms of value-ladenness need to be distinguished. It is widely acknowledged that epistemic values, i.e. values indicative of knowledge or truth, such as predictive accuracy, internal coherence, and explanatory power, play a variety of legitimate roles in science (Douglas 2009; Elliott and McKaughan 2014; Rooney 2017).

It is more controversial how non-epistemic values (henceforth only "values"), such as ethical and political values, can legitimately figure in science.[4] In particular, it is controversial whether they figure in the internal stage of science (Douglas 2009) and in "doing science" (Elliott 2022a), which concern the justification of scientific hypotheses and involve processes such as study design, data analysis, and the interpretation of results. By contrast, the external stage involves "steering science" (e.g., decisions about what topic to study), "using science" (e.g., for policy-making), and "managing science" (e.g., setting codes of conduct for scientists) (Elliott 2022a, sect. 2.3). It is obvious that values appear in the external stage.

While the boundary between the internal and the external stage is not strict (for instance, because steering research into certain directions is related to questions of research design (Elliott 2022a, p. 9)), this is, nevertheless, a viable distinction. It seems obvious that, for all sciences, values affect and justify[5] (e.g.) which topics are studied, which codes of scientific conduct are adopted, and which decisions are made based on scientific results. Thus, that the same applies to AIA is not surprising.

3.2 Value-ladenness and strong value-ladenness

[4] However, several authors have questioned the tenability of the distinction between epistemic and non-epistemic values (Longino 1996; Rooney 2017). In this paper, I do not commit to a view on these critiques.
[5] Plausibly, the external stage is value-laden in all four senses of the term identified by Ward (2021).

If (non-epistemic) values are involved in the internal stage of a science, then this entails that the respective science is *value-laden*. Some philosophers have influentially argued that many sciences are value-laden in this sense (e.g. Douglas 2000; Longino 1996). I will presuppose the soundness of these arguments here, so that I can explore their implications for AIA. I hold, in addition, that some sciences, or subdisciplines, are more value-laden than others. Call these *strongly value-laden*. This idea has intuitive appeal:

> [I]t is worth noting that values play a role in all fields of science, but those roles are even more prevalent in some fields of science than in others. Value-laden judgements are particularly pronounced in fields where scientists' conclusions have fairly direct implications for social decision-making (e.g., agricultural research, toxicology, environmental science, and many areas of the biomedical and social sciences). In contrast, ethical and social values have fewer obvious roles to play in deciding how to model, interpret, and categorise phenomena in more theoretical areas of science (e.g., chemistry and physics) that are fairly disconnected from social decision-making (Elliott 2022a, p. 14).

This quote suggests a criterion for strong value-ladenness: a science is particularly value-laden if its results have especially direct and important relevance for social decision-making. This claim can be supported in recourse to Douglas' (2000, 2009) *inductive risk argument*. In a nutshell, Douglas argues that – when making inductive inferences – researchers face a risk of error. This *inductive risk* may be distributed asymmetrically. According to the argument, choices of methodology, data characterization, interpretation of results etc. affect whether scientists are more likely to make a false-positive or a false-negative judgement. This is why, according to Douglas, scientists should consider whether false-positives or false-negatives are more harmful and thus rely on values to decide which methodological choices to make and how much evidence to demand in order to accept or reject hypotheses. If the inductive risk argument captures one central reason why (induction-based) sciences are value-laden, then it is plausible that they are strongly value-laden if (i) they involve high inductive risk, i.e., high chances of error, and (ii) errors have particularly significant and immediate effects on society.

A second criterion of strong value-ladenness is suggested by Alexandrova's (2017) analysis of mixed claims in wellbeing science. Following Alexandrova, I can call a claim mixed iff it is an empirical claim about some causal or statistical relation and "[a]t least one of the variables in this claim is defined in a way that presupposes a moral, prudential, or political value judgement about the nature of this variable" (Alexandrova 2017, p. 82). Mixed claims contain concepts which combine empirical and evaluative elements, such as wellbeing, efficiency, aggression, or rape. Moreover, Alexandrova argues that the role of mixed claims is unique: mixed claims are distinct from other forms of value-ladenness in science.

I hold that the presence of mixed claims is a viable second criterion for strong value-ladenness. First, not all sciences contain mixed claims (while all sciences may be value-laden), so mixed claims carve out a special class of sciences. Second, mixed claims entail a particularly intimate way in which sciences can be value-laden. Due to mixed claims, it is not only the case that scientists ought to consider values in their scientific decisions (as the inductive risk argument supposes), but that the judgements scientists make also themselves express values. In other words, given mixed claims, value-ladenness is not derivative of scientist's decisions, it is contained in scientific outputs – statements, models, or theories – themselves.

I do not claim that these criteria are exhaustive: there may be other criteria for determining the degree to which a science is value-laden. Moreover, I do not make any substantive ontological claim to the effect that value-ladenness, itself, has degrees: I merely rely on the plausible idea that values are more or less prevalent and influential in different sciences.

To summarize, a research field is strongly value-laden if it contains much uncertainty, has especially high relevance for social decision-making, and centrally involves mixed claims. I will now show that AIA is strongly value-laden in this sense. Then, I will consider challenges to my argument.

3.3 Values in AIA

First, alignment claims presuppose value judgements about whether a system's behavior is desirable. Thus, alignment claims are mixed claims. Such claims are obviously central to AIA. For example, Bai, Jones et al. report key results as follows (2022): "Our alignment interventions actually enhance the capabilities of large models, and can easily be combined with training for specialized skills (such as coding or summarization) without any degradation in alignment or performance." The truth of claims about when a system is aligned may, for example, be dependent on controversial value assumptions when people disagree about when RLHF has accomplished to make a language model helpful, harmless, and honest. Assessments of whether certain text outputs are racially biased, and how harmful this is, depend on ethical and political values. Also, value conflicts suggest conflicting judgements on trade-offs between helpfulness, harmlessness, and honesty. So, alignment judgements depend on values.

This value sensitivity extends to general judgements about the properties of alignment (as the one cited last paragraph). One cause of this is that some ethical conceptions may be easier to align to than others. For example, some might think that helpfulness of answers is less correlated with what users rate as helpful than assumed. They might think, e.g., that longer,

sophisticated and more nuanced answers are often more helpful than users rate them to be because they are rewarding, but take time and concentration to process. If we understand helpfulness this way, then – using RLHF – systems may turn out to be harder to align than assumed. This might undermine claims about the conditions for aligning LLMs. So, which causal generalizations we take to hold of alignment depends on which value judgements we accept.

A potential reply is that the previous discussion conflates technical and ethical alignment. According to this reply, AIA proper does not involve mixed claims: it only concerns *how* to make systems aligned, given some antecedently specified ethical notion of alignment. Which ethical notion to choose is outside of the scope of AIA. This reply is analogous to a typical move in debates on values in science. According to this move, science can be understood as making conditional claims: claims about what is true, *if* certain values are assumed (Betz 2013, 2017).

It is disputed whether this strategy is tenable as a normative recommendation for scientists (Alexandrova 2017, chapter 4; Elliott 2022a). However, it is clearer that this reply is not capturing current practice in AIA: it is not the case that AIA researchers are always presenting their results as contingent on certain value assumptions. Of course, empirical operationalizations of ethically laden notions (such as "safety" or "harmlessness") have to be used in AIA research, and are used, to make its research questions empirically tractable. Nevertheless, papers typically lack an explicit discussion of the ethical assumptions these operationalizations, and consequently the papers' overall research results, are justified by and sensitive to, and they do not elaborate on the consequences alternative ethical assumptions would have. Similarly, they often do not distinguish narrow technical alignment questions from the broader ethical issues involved and do not mark the latter explicitly as ethical (or normative or something similar).

A comprehensive review of the AIA literature would be needed to confirm empirically how frequently AIA researchers present their claims as conditional on their value assumptions. However, my claims are correct, at least, for most or all of the studies cited above. For example, none of the abstracts of the empirical AIA studies cited in 2.2 explicitly identifies any of its assumptions as "ethical", "normative" or the like, or classifies its conclusions as dependent on such assumptions. This is the case even though many of these studies make mixed claims and have high inductive risk, as I have argued.

Explicit discussions of ethical assumptions which might satisfy the standards described above also tend to be absent from the main text. Bai, Jones et al.'s (2022) canonical paper on

RLHF is an interesting case here. They explicitly note at the beginning of their paper that they do not define helpfulness and harmlessness and leave it open to the judgement of the crowdworkers to assess LLM outputs in these terms. Moreover, they contend that "ethical, legal, and cultural expertise" (section 7.2) is helpful to ultimately determine which AI-behavior is preferrable, thus acknowledging the value-ladenness of AI alignment.

At the same time, they say that using judgements of crowdworkers to define helpfulness and harmlessness "was sufficient for our exploration of 'technical alignment'" (section 7.2). Thus, they seem to presuppose that ethical questions can be set aside in empirically driven AIA. Perhaps for this reason, the value assumption that helpfulness and harmless can be operationalized via crowdworker judgements does not play a role in the rest of the discussion of their results and they do not report their main claims as conditional on certain value assumptions. However, if my arguments in this sub-section are correct, then even technical alignment research is not value-free. So, even if one thinks this conditionalization strategy is promising, the examples of current AIA cited above suggest that AIA currently often presupposes value claims.

Second, when making inferences, AIA researchers face inductive risk. For instance, Hubinger et al.'s study (2024) reveals limitations of currently dominant alignment techniques. Given that AIA is a young research field without a general consensus about methods, concepts and foundational theoretical assumptions, pervasive uncertainty, and thus risk of error, is unavoidable. This inductive risk is sometimes distributed asymmetrically. That is, taking systems to be aligned when they are not may, in many cases, be worse than the reverse (though the opposite is also conceivable). Burns et al. (2023) even justify their methodology in exploring alignment of super-human AI systems by considerations of inductive risk saying that "[g]iven the stakes, we need to establish *extremely high* reliability in the alignment of these systems ahead of time" [my emphasis].

Moreover, some results of AIA feed directly into social decision-making. Which systems are deemed "aligned" influences which systems are deployed by AI companies for public use (Shevlane et al. 2023). Similarly, regulation of AI may likely be sensitive to the results of AIA and their framing: if certain classes of models are deemed harder to align, for example, they may be more heavily regulated. Also, in some eyes, the risks at stake are very big. Some ethicists think that language models may cause severe amounts of harm by influencing public discourse or by perpetuating racist, sexist and other biases (Bender et al. 2021). Other researchers emphasize catastrophic, e.g. through increasing the accessibility of bioweapons (Soice et al. 2023), or even extinction risks from misaligned AI (Center for AI Safety 2023).

No matter how one stands on these particular concerns: the fact that researchers, citizens, and policymakers commonly have views according to which the stakes – potential benefits and harms – of AI development and deployment are very high, in conjunction with the observation that alignment judgements influence the estimation of these risks and benefits, entails that some errors in alignment judgements matter very much, morally speaking. If so, then AIA meets the first criterion for strong value-ladenness: its results are especially relevant for social decision-making.

In line with Betz (2013), one might object that AIA researchers could minimize inductive risk by only making conditional statements, where all uncertain empirical and value assumptions on which the truth of a claim depends appear in the antecedent of the conditional. My response is the same as to the previous conditionalization challenge: even if this strategy is feasible (which is contested), it is not currently being consistently applied in AIA. While I have argued this previously for ethical assumptions, an advocate of Betz' strategy needs, moreover, to demand that empirical uncertainty is also removed by conditionalizing on uncertain empirical assumptions. I don't see any reason to think AIA currently satisfies this demand. Thus, currently, AIA makes statements which face serious inductive risks.

3.4 Challenges and contextualization

I have argued that high and immediate social relevance and a central role of mixed claims are jointly sufficient criteria for a scientific discipline to be strongly value-laden, and that AIA meets these criteria. I will now address an objection and provide contextualization. First, Peters (2023) questions two claims Alexandrova (2017) makes about mixed claims in the science of wellbeing: 1. The presence of mixed claims entails that a science is value-laden, and 2. The value-ladenness of mixed claims is distinct from other kinds of value-ladenness of science. If these views are false, then this undermines my view that mixed claims can serve as a criterion of strong value-ladenness.

However, Peters relies on the same kind of conditionalization argument we have already discussed. So, my response to Peters' argument is similar to the previous objection that AIA may be taken to be only focused on technical alignment, rather than a mixed alignment notion. Peters' argument shows that, in principle, it is possible to conditionalize on the value judgements made in AIA and to limit oneself to only making empirical claims, formulated as conditionals with value assumptions in the antecedent. If so, the discussion of the requisite value assumptions is relegated outside of the internal stage of AIA. In this case, the presence of mixed claims in AIA does not even entail any kind of value-ladenness, and the evaluative role

of mixed claims is reducible to the uncontroversial role of values as agenda setters, i.e. influencing the choice of research topic.

I take this to mean: if successful, Peters' argument shows that mixed claims do not have to figure in the internal stage of AIA and, if they do not, they function as mere agenda-setters. Yet, in current AIA practice, such a conditionalization strategy is not consistently pursued. To be clear, ethical assumptions have to be operationalized somehow, so – in AIA – ethical notions like "helpfulness, harmless, and honesty" are typically explicitly identified with the outcomes of some empirical process, e.g. user ratings. However, this arguably falls short of the demand of conditionalization on ethical assumptions that the previous researchers identify. This demand arguably also requires that ethical assumptions are identified as such, that the results of a study (most centrally, in the abstract and the "results" section of a paper) are explicitly presented as conditional on these assumptions, and that the possibility of contrasting reasonable ethical assumptions is noted and its significance mentioned. Since this demand is not fulfilled, AIA – as it is currently practiced – contains mixed claims in its internal stage which entails strong value-ladenness.

To further situate my claim in the literature, I note that several authors have already provided arguments according to which the development and design of machine learning (ML) systems is value-laden (Biddle 2022, 2023; Johnson forthcoming; Nyrup 2022). While these authors do not talk about strong value-ladenness in particular, there appear to be reasons to think that their claims entail that ML research generally is strongly value-laden. While this would support my claims about AIA, it would also make them less distinctive. However, I don't think that the distinctiveness of AIA with respect to value-ladenness is threatened, since strong value-ladenness is less ubiquitous in ML overall than in AIA specifically and, moreover, the value-ladenness in ML sometimes depends on AIA.

The first reason to think that ML research is strongly value-laden is that all of these authors, although framed in different ways, employ the inductive risk argument (although they don't all rely only on the inductive risk argument). Moreover, they cite ML systems, such as algorithms designed to inform parole decisions (Biddle 2022; Johnson forthcoming), which are poised to directly inform socially significant decisions. Second, while ML research does not explicitly center on a mixed term, one may argue that decisions regarding socially significant systems nevertheless often implicitly, and sometimes explicitly, invoke discussions about mixed terms such as "safety" or "discrimination".

In response, I hold that these arguments show that some parts of ML are strongly value-laden, although many might not be. The projects within ML research which pose especially

socially significant inductive risks may loosely correspond to the domains the EU AI Act considers "high risk" (or unacceptable risk), that is, research on systems poised to be used in education, critical infrastructure, law enforcement etc. However, much of ML research is not directly relevant for applications in such risky domains, so it will not pass my second criterion of strong value-ladenness.

Moreover, if ML research is strongly value-laden, this is often the case when it blends into AIA, since the borders of the latter are not sharp. For instance, research involving questions about how to build decision-making systems such that they make "fair" and "unbiased" decisions plausibly fits the definition of alignment: it is about how to build AI systems whose decision-making corresponds to certain ethical conceptions. The same is true of mixed terms. Many ML projects don't involve reasoning about mixed terms (e.g., building systems with high accuracy in image recognition). Moreover, if they do, these are often mixed terms which also appear in AIA, such as "safety" (the goal of alignment is motivated by the aim to make systems safe). So, while some domains of ML are strongly value-laden, others are not. The ones which are strongly value-laden often overlap with AIA, thus positioning AIA as an especially important case of value-ladenness, which explains value-ladenness in some other areas of ML.

Of course, without precisely delineating the extension of the term "AIA", as I use it here, I cannot claim that all areas of AIA are value-laden. Moreover, some research which may be classified as AIA concerns low-stakes research on well-understood systems and may thus have low inductive risk (e.g. foundational interpretability research on very simple systems). So, it is plausible that not all of AIA is value-laden. However, since "alignment" and cognate terms like "safety" are mixed terms which are central to AIA and since AIA is more directly connected to ethical issues than much of ML, it seems plausible that – compared to ML research generally – much of AIA is value-laden, and even strongly value-laden.

In this section, I have argued that AIA is strongly value-laden. In the next section, I turn to the question how the value-ladenness of AIA should be managed.

4. Managing values in AI alignment research

4.1 Threats of strong value-ladenness

In this sub-section, I claim that values in AIA are currently not responsibly handled. In particular, their management lacks two features: First, those values are often not made explicit. As previously suggested, researchers rarely discuss how to conceive of alignment or explicate which value-laden decisions they made in the face of inductive risk. If the role of values in AIA is not transparent, then AIA may – misleadingly – appear as a value-free enterprise. Second,

and relatedly, AIA does not use any systematic method for determining which value-judgements are appropriate. Empirical research employs methods conducive to finding truths about the empirical world. Similarly, AIA should employ methods which are – though fallible – conducive to making value assumptions which are true, politically legitimate, or in some other way warranted and appropriate.[6]

These two shortcomings create risks. The first risk is that values may compromise AIA's ability to produce knowledge. If the value assumptions in AIA are false, then the claims depending on them may be false. If the value assumptions in AIA lack any justification, then the claims depending on them may be unjustified. For instance, if systems are aligned via RLHF, then assessments of whether they are aligned may involve values and thus be false or unjustified, if we lack good reasons for the requisite value assumptions (such as the assumption that judgements of crowdworkers faithfully track helpfulness and harmlessness of LLM outputs).

The second risk is that AIA may be ethically useless or even harmful, if it rests on value assumptions which are false or otherwise illegitimate. For example, if AIA researchers weigh inductive risks wrongly, AIA may induce a sense of false security that systems are safe which then suddenly gives rise to catastrophic failure (for instance via "deceptive alignment" (Dung 2023, section 5.3; Park et al. 2023)). In addition, claims about when language models are aligned may mask value beliefs of researchers about what LLMs should (not) be allowed to say as scientific statements, and thus might cause them to illegitimately and inadvertently impose their values on the population (Alexandrova 2017, section 4.4). Moreover, some researchers have criticized the contention, which seems to be implicitly presupposed by many AIA projects, that it is desirable to create super-human AI, if it is aligned (in a sense not always further specified) (Friederich 2023; Sparrow 2023).

While some of these dangers may obtain even with value-ladenness which is not strong, they are particularly pressing in cases of strong value-ladenness. In this case, scientific hypotheses themselves contain value presuppositions, making the former's epistemic status dependent on the latter. Furthermore, if strong value-ladenness obtains, then researchers have strong reasons to take the ethical impacts of inductive errors into account. If these errors have more negligible or indirect impacts, which are hard to anticipate anyway, then it may be wise to set them aside.

[6] I intend to stay neutral on metaethical questions regarding the semantics, epistemology, or metaphysics of values.

Hence, strong value-ladenness raises the risk that AIA is epistemically or ethically harmful. In the next section, I will make recommendations for mitigating these risks.

4.2 Proposals for value management in AIA

In light of these risks, I will make several recommendations. In general, approaches for the management of values in science can be distinguished in terms of at least three features. First, *transparency approaches* are about making the influence of values explicit, while *value choice approaches* concern how particular values should be chosen. Though transparency plausibly has some downsides (Nguyen 2022), it is almost always regarded as crucial to the management of values in science (Douglas 2009; Elliott 2022b). Nyrup (2022, p. 1056) holds that transparency requires the following:

> (a) Openness: communicating (or making accessible) the relevant information to the right audience.
> (b) Comprehensibility: ensuring the audience can understand and use the information. […]
> (c) Explicitness: ensuring the communicator is aware of and able to articulate the information in the right way, i.e., so that it achieves (a) and (b).

Moreover, he distinguishes two types of value transparency: *retrospective value transparency* concerns the values which motivated and influenced a given decision, and *prospective value transparency* concerns the impacts of decisions on things we value; and the values which would justify these decisions.

To achieve these kinds of value transparency, researchers should not exaggerate the extent to which AIA is value neutral; instead, they should formulate its value presuppositions explicitly. Crucially, they should emphasize specifically where value presuppositions made in AIA are potentially controversial. Relatedly, AIA researchers should analyze which value assumptions their research results are sensitive to. If values pervasively affect many steps of the scientific process, it may be impossible to communicate all of these assumptions (Alexandrova 2017, section 4.3). However, value assumptions such as the ones I used here as examples, e.g. the operationalization of chatbot alignment as helpfulness, harmlessness, and honesty, which are particularly important for the research results and which are sufficiently salient should be made explicit.

This contrasts with common talk about values in AIA which is either not critical and aimed at enhancing the reputation of a commercial AI firm (e.g. Leike et al. 2022), or too general to uncover important value disagreements. As an example of the latter, consider the RICE principles which suggest that AIA should aim at robust, interpretable, controllable, and ethical AI (Ji et al. 2023). As an example, "Ethicality refers to a system's unwavering commitment to uphold universally acknowledged norms and values within its decision-making

and actions“ (Ji et al. 2023, section 1.1.2). Such principles are too broad to shed light on actual value disagreements about AIA, e.g., about the risks of describing alignment techniques as successful when they may break down when applied to hypothetical future systems with superhuman capacities. So, the envisioned value transparency is more specific. Overall, value transparency allows the recipients of scientific research to make informed decisions about whether they accept the presupposed values, and thus the scientific results contingent on them.

However, transparency is not enough. If value presuppositions are unreasonable or illegitimate, then value transparency is not going to change this: it will just make it transparent. Value choice approaches concern what value presuppositions to make. Second, Schroeder (2022a) distinguishes *political* and *ethical* approaches to value choice. Political approaches employ the normative standard of democratic political legitimacy: as a result, they hold that the values chosen by scientists should in some way be grounded in public values (Schroeder 2021). By contrast, ethical approaches employ the normative standard of substantive ethical correctness: as a result, they hold that ethical reasoning or principles determine which values scientists should choose.

While I cannot resolve this disagreement here, I submit that a combination of both approaches may be optimal. AIA researchers should critically scrutinize their value assumptions. Since value judgements may be unconscious and thus not always accessible to one’s own scrutiny, researchers should also examine the value assumptions made by others. Ideally, this discussion should happen with the same rigor and sophistication with which technical research is pursued. But complementarily, actors in AIA should encourage critical discussion and scrutiny of AIA’s value assumptions beyond the confines of their discipline. This can take several forms. To some extent, these conversations may happen in public fora in the day-to-day lives of ordinary citizens. However, since such discussions can also be counterproductive, partially due to distorting incentives of media companies such as Facebook, X, or newspapers, alternatives may be preferable. For instance, Citizens’ panels may be used. A Citizens’ panel is a large, demographically representative group of citizens which is first informed by experts and then jointly deliberates on some important normative questions, e.g. about public policy. Citizens’ panels could examine which value assumptions are appropriate for AIA. Finally, experts from other relevant discipline, such as ethics, should be encouraged to scrutinize the values within AIA. Lay-user studies, employed in the field of explainable AI to confirm that AI systems are tailored to their users’ needs and values (Rong et al. 2023), are an example of fruitful engagement between AIA and the public.

In particular, consulting the views of members of the public mitigates the danger that researchers illegitimately impose their values on the public (Alexandrova 2017, chapter 4). It is plausible that members of the public should be involved in value-based scientific decisions which may have profound impacts on society, and thus on them. At the same time, popular opinion or the verdicts of Citizens' panels should not be treated as the sole normative standard AIA is subject to, but as one factor to consider (for one class of cases where scientists should not defer to the public, see Schroeder 2022b). Since value and scientific questions are entangled, AIA researchers may have some special expertise relevant to assessing values in AIA (although they also lack other relevant forms of expertise, e.g. of ethicists).[7] Moreover, many people became AIA researchers because of their ethical beliefs, e.g. concern about AI existential risk. In a liberal society, there should be some space for private citizens to try to improve the world, no matter whether the general public agrees with them. Hence, it is plausibly legitimate – within some bounds – to become an AIA researcher who works based on idiosyncratic personal value assumptions. So, AIA's practices should be informed and modestly constrained, but not handcuffed by public (or expert) opinion.

Moreover, in cases in which finding the appropriate value assumption is difficult, AIA researchers can sometimes communicate their results as conditionals, relative to these assumptions (Betz 2013). It may be impossible or infeasible to apply this procedure to all value assumptions in AIA. For if values are present throughout the research process in AIA and one understands all AIA results as conditionals, one may quickly end up with an intricate web of conditional relations, which it is not feasible to handle. However, some hard-to-decide and important value assumptions should be treated, to some degree, via conditionalization: researchers should explicitly state that their claims as conditionals, with these assumptions in the antecedent. However, since value judgements are in turn necessary to determine which alternatives to consider, the need for some value judgements appears *inescapable*. Combined, using the different procedures I have outlined in value choice would serve to increase AIA's democratic legitimacy as well as the chance that AIA's value presuppositions are substantively ethically true or reasonable.

A third feature is whether values are managed on the level of *individual scientists* or research groups or on the level of the *scientific community*. For instance, Longino (1990, 2001) maintains that the objectivity of science, in face of its value-ladenness, does not depend on the decisions of individual scientists, but on whether the scientific community is structured such to

[7] Alexandrova (2017, chapter 4) makes this claim about wellbeing researchers.

adequately facilitate critical reflection. I believe that individual and community approaches both have a role to play. On the level of the community, I propose that diversifying the funding situation within AIA could contribute to the appropriate management of values.

Currently, leading AIA is mostly performed by a small, homogenous group of profit-oriented companies – the same companies which also have a leading role in increasing AI capabilities (chiefly, Google, Microsoft via OpenAI, and Anthropic). This is problematic because profit-oriented companies have strong incentives to sacrifice ethical motives if they collide with their desire for profit. Thus, there is a permanent risk that these companies practice "alignment washing": continue to work on AIA to enhance their public image, but without considering strategies which may be good for alignment and safety but bad for their commercial interests. Profit motives may even incentivize these companies to not only conduct ineffectual AIA, but also to distort the research priorities and processes of the field at large (comparable to, for instance, Tobacco companies which selectively funded researchers claiming that smoking does not cause cancer).

However, it is nevertheless desirable that these companies work on AIA. Arguably, progress on AI alignment is badly needed. Moreover, if one would stop these companies from pursuing AIA, they would likely use the spare resources to just build more capable AI systems – without special concern for ethics and safety. This is hardly a better outcome. However, to ameliorate the current situation, the amount of *public* funding for AIA could be increased. Arguably, AIA which is conducted by organizations which are supported by public funding, chiefly universities, would be less susceptible to distorting commercial incentives. Even if this is not the case, non-profit actors and actors from different countries and cultures may be inclined to make different value presuppositions than Silicon Valley based for-profit actors, thus increasing the diversity of value presuppositions in AIA and making critical scrutiny of value presuppositions from within the AIA community possible (for a possible argument for such value diversity, see Thoma 2023). So, even if no specific institutional context for AIA research can be shown to be superior, more diversity is desirable, nevertheless. Consequently, increasing the amount of public funding for AIA would lead to a more balanced, robust, and diverse AIA landscape.

I hold that values in AIA do not undermine its epistemic and ethical aspirations *if* my suggestions are implemented. However, since current AIA falls short of this, its reliance on values currently threatens its epistemic and ethical success.

5. Conclusion

In this paper, I have argued for several claims. First, sciences differ in the degree to which values influence them. Second, given this distinction, AIA is particularly – strongly – value-laden. Third, this influence of values is currently not managed appropriately and thus threatens the epistemic integrity and ethical beneficence of AIA. Fourth, in response AIA should strive to achieve value transparency, critical scrutiny from inside and outside the discipline – involving the public –, and to empower actors without strong commercial interests to assume a more important role in AIA. An overarching lesson is that the field of empirical AI alignment research would benefit from the scrutiny of philosophers of science.

Acknowledgements

I thank Uwe Peters for helpful comments and discussion.

References

Alexandrova, A. (2017). *A Philosophy for the Science of Well-Being* (Vol. 1). Oxford University Press. https://doi.org/10.1093/oso/9780199300518.001.0001

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., et al. (2022, April 12). Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. arXiv. https://doi.org/10.48550/arXiv.2204.05862

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., et al. (2022, December 15). Constitutional AI: Harmlessness from AI Feedback. arXiv. https://doi.org/10.48550/arXiv.2212.08073

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Betz, G. (2013). In defence of the value free ideal. *European Journal for Philosophy of Science*, *3*(2), 207–220. https://doi.org/10.1007/s13194-012-0062-x

Betz, G. (2017). Why the Argument from Inductive Risk Doesn't Justify Incorporating Non-Epistemic Values in Scientific Reasoning. In K. C. Elliott & D. Steel (Eds.), *Current Controversies in Values and Science* (pp. 94–110). London: Routledge.

Biddle, J. B. (2022). On Predicting Recidivism: Epistemic Risk, Tradeoffs, and Values in Machine Learning. *Canadian Journal of Philosophy*, *52*(3), 321–341. https://doi.org/10.1017/can.2020.27

Biddle, J. B. (2023). Values in Artificial Intelligence Systems. In *Technology Ethics*. Routledge.

Bostrom, N. (2014). *Superintelligence. Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., et al. (2023, March 22). Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv. https://doi.org/10.48550/arXiv.2303.12712

Burns, C., Izmailov, P., Kirchner, J. H., Baker, B., Gao, L., Aschenbrenner, L., et al. (2023, December 14). Weak-to-Strong Generalization: Eliciting Strong Capabilities With Weak Supervision. arXiv. https://doi.org/10.48550/arXiv.2312.09390

Butlin, P. (2023). Reinforcement learning and artificial agency. *Mind & Language*, mila.12458.

https://doi.org/10.1111/mila.12458
Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., et al. (2023, July 27). Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. arXiv. https://doi.org/10.48550/arXiv.2307.15217
Center for AI Safety. (2023). Statement on AI Risk. https://www.safe.ai/statement-on-ai-risk. Accessed 20 June 2023
Christian, B. (2020). *The Alignment Problem: Machine Learning and Human Values*. W. W. Norton & Co.
Cotra, A. (2021, September 21). Why AI alignment could be hard with modern deep learning. *Cold Takes*. https://www.cold-takes.com/why-ai-alignment-could-be-hard-with-modern-deep-learning/. Accessed 15 January 2023
Douglas, H. (2000). Inductive Risk and Values in Science. *Philosophy of Science*, *67*(4), 559–579. https://doi.org/10.1086/392855
Douglas, H. (2009). *Science, Policy, and the Value-Free Ideal*. Pittsburgh: University of Pittsburgh Press.
Dung, L. (2023). Current cases of AI misalignment and their implications for future risks. *Synthese*, *202*(5), 138. https://doi.org/10.1007/s11229-023-04367-0
Dung, L. (2024a). The argument for near-term human disempowerment through AI. *AI & SOCIETY*. https://doi.org/10.1007/s00146-024-01930-2
Dung, L. (2024b). Understanding Artificial Agency. *The Philosophical Quarterly*, pqae010. https://doi.org/10.1093/pq/pqae010
Elliott, K. C. (2022a). *Values in Science* (1st ed.). Cambridge University Press. https://doi.org/10.1017/9781009052597
Elliott, K. C. (2022b). A Taxonomy of Transparency in Science. *Canadian Journal of Philosophy*, *52*(3), 342–355. https://doi.org/10.1017/can.2020.21
Elliott, K. C., & McKaughan, D. J. (2014). Nonepistemic Values and the Multiple Goals of Science. *Philosophy of Science*, *81*(1), 1–21. https://doi.org/10.1086/674345
Friederich, S. (2023). Symbiosis, not alignment, as the goal for liberal democracies in the transition to artificial general intelligence. *AI and Ethics*. https://doi.org/10.1007/s43681-023-00268-7
Gabriel, I. (2020). Artificial Intelligence, Values, and Alignment. *Minds and Machines*, *30*(3), 411–437. https://doi.org/10.1007/s11023-020-09539-2
Hubinger, E. (2020, December 4). An overview of 11 proposals for building safe advanced AI. arXiv. https://doi.org/10.48550/arXiv.2012.07532
Hubinger, E. (2021). How do we become confident in the safety of a machine learning system? https://www.alignmentforum.org/posts/FDJnZt8Ks2djouQTZ/how-do-we-become-confident-in-the-safety-of-a-machine. Accessed 10 August 2023
Hubinger, E., Denison, C., Mu, J., Lambert, M., Tong, M., MacDiarmid, M., et al. (2024, January 10). Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. arXiv. http://arxiv.org/abs/2401.05566. Accessed 12 January 2024
Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., et al. (2023, November 1). AI Alignment: A Comprehensive Survey. arXiv. https://doi.org/10.48550/arXiv.2310.19852
Johnson, G. M. (forthcoming). Are Algorithms Value-Free? Feminist Theoretical Virtues in Machine Learning. *Journal Moral Philosophy*, 1–35. https://doi.org/10.1163/17455243-20234372
Kirchner, J. H., Smith, L., Thibodeau, J., McDonell, K., & Reynolds, L. (2022, June 6). Researching Alignment Research: Unsupervised Analysis. arXiv. http://arxiv.org/abs/2206.02841. Accessed 18 January 2024
Korbak, T., Shi, K., Chen, A., Bhalerao, R., Buckley, C. L., Phang, J., et al. (2023, June 14). Pretraining Language Models with Human Preferences. arXiv. https://doi.org/10.48550/arXiv.2302.08582

Kuhn, T. S. (1962). *The structure of scientific revolutions*. University of Chicago Press: Chicago.
Langosco, L., Koch, J., Sharkey, L., Pfau, J., Orseau, L., & Krueger, D. (2023, January 9). Goal Misgeneralization in Deep Reinforcement Learning. arXiv. https://doi.org/10.48550/arXiv.2105.14111
Leike, J., Schulman, J., & Wu, J. (2022). Our approach to alignment research. *openai.com*. https://openai.com/blog/our-approach-to-alignment-research. Accessed 11 August 2023
Longino, H. E. (1990). *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton: Princeton University Press. https://press.princeton.edu/books/paperback/9780691020518/science-as-social-knowledge. Accessed 14 February 2024
Longino, H. E. (1996). Cognitive and Non-Cognitive Values in Science: Rethinking the Dichotomy. In L. H. Nelson & J. Nelson (Eds.), *Feminism, Science, and the Philosophy of Science* (pp. 39–58). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-009-1742-2_3
Longino, H. E. (2001). *The Fate of Knowledge*. Princeton University Press.
Nguyen, C. T. (2022). Transparency is Surveillance. *Philosophy and Phenomenological Research*, *105*(2), 331–361. https://doi.org/10.1111/phpr.12823
Nyrup, R. (2022). The Limits of Value Transparency in Machine Learning. *Philosophy of Science*, *89*(5), 1054–1064. https://doi.org/10.1017/psa.2022.61
OpenAI. (2016, December 22). Faulty Reward Functions in the Wild. https://openai.com/blog/faulty-reward-functions/. Accessed 14 January 2023
Pacchiardi, L., Chan, A. J., Mindermann, S., Moscovitz, I., Pan, A. Y., Gal, Y., et al. (2023, September 26). How to Catch an AI Liar: Lie Detection in Black-Box LLMs by Asking Unrelated Questions. arXiv. https://doi.org/10.48550/arXiv.2309.15840
Pan, A., Bhatia, K., & Steinhardt, J. (2022, February 14). The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models. arXiv. https://doi.org/10.48550/arXiv.2201.03544
Park, P. S., Goldstein, S., O'Gara, A., Chen, M., & Hendrycks, D. (2023, August 28). AI Deception: A Survey of Examples, Risks, and Potential Solutions. arXiv. https://doi.org/10.48550/arXiv.2308.14752
Peters, U. (2023). Values in science: assessing the case for mixed claims. *Inquiry: An Interdisciplinary Journal of Philosophy*, *66*(6), 965–976. https://doi.org/10.1080/0020174X.2020.1712235
Rong, Y., Leemann, T., Nguyen, T., Fiedler, L., Qian, P., Unhelkar, V., et al. (2023, December 19). Towards Human-centered Explainable AI: A Survey of User Studies for Model Explanations. arXiv. https://doi.org/10.48550/arXiv.2210.11584
Rooney, P. (2017). The Borderlands between Epistemic and Non-Epistemic Values. In K. C. Elliott & D. Steel (Eds.), *Current Controversies in Values and Science* (1st ed., pp. 31–45). New York: Routledge. https://doi.org/10.4324/9781315639420
Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.
Schlarmann, C., & Hein, M. (2023, August 21). On the Adversarial Robustness of Multi-Modal Foundation Models. arXiv. https://doi.org/10.48550/arXiv.2308.10741
Schroeder, S. A. (2021). Democratic Values: A Better Foundation for Public Trust in Science. *The British Journal for the Philosophy of Science*, *72*(2), 545–562. https://doi.org/10.1093/bjps/axz023
Schroeder, S. A. (2022a). Thinking about Values in Science: Ethical versus Political Approaches. *Canadian Journal of Philosophy*, *52*(3), 246–255. https://doi.org/10.1017/can.2020.41
Schroeder, S. A. (2022b). The Limits of Democratizing Science: When Scientists Should Ignore the Public. *Philosophy of Science*, *89*(5), 1034–1043. https://doi.org/10.1017/psa.2022.54
Shevlane, T., Farquhar, S., Garfinkel, B., Phuong, M., Whittlestone, J., Leung, J., et al. (2023, May 24). Model evaluation for extreme risks. arXiv. https://doi.org/10.48550/arXiv.2305.15324

Soice, E. H., Rocha, R., Cordova, K., Specter, M., & Esvelt, K. M. (2023, June 6). Can large language models democratize access to dual-use biotechnology? arXiv. https://doi.org/10.48550/arXiv.2306.03809

Sparrow, R. (2023). Friendly AI will still be our master. Or, why we should not want to be the pets of super-intelligent computers. *AI & SOCIETY*. https://doi.org/10.1007/s00146-023-01698-x

Thoma, J. (2023). Social Science, Policy and Democracy. *Philosophy and Public Affairs*, *52*(1), 5–41. https://doi.org/10.1111/papa.12250

Turner, A. (2022). Inner and outer alignment decompose one hard problem into two extremely hard problems — AI Alignment Forum. https://www.alignmentforum.org/posts/gHefoxiznGfsbiAu9/inner-and-outer-alignment-decompose-one-hard-problem-into. Accessed 22 January 2024

Ward, Z. B. (2021). On value-laden science. *Studies in History and Philosophy of Science Part A*, *85*, 54–62. https://doi.org/10.1016/j.shpsa.2020.09.006

Yudkowsky, E. (2016). *The AI alignment problem: Why it is hard, and where to start*. Presented at the Symbolic Systems Distinguished Speaker. https://intelligence.org/stanford-talk/. Accessed 18 January 2024

Ze Shen, C. (2022). A newcomer's guide to the technical AI safety field. https://www.alignmentforum.org/posts/5rsa37pBjo4Cf9fkE/a-newcomer-s-guide-to-the-technical-ai-safety-field. Accessed 18 January 2024

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., et al. (2023, October 2). Representation Engineering: A Top-Down Approach to AI Transparency. arXiv. https://doi.org/10.48550/arXiv.2310.01405

Zou, A., Wang, Z., Kolter, J. Z., & Fredrikson, M. (2023, July 27). Universal and Transferable Adversarial Attacks on Aligned Language Models. arXiv. https://doi.org/10.48550/arXiv.2307.15043