AI and Ethics
https://doi.org/10.1007/s43681-025-00703-x

RESEARCH

Metaethical perspectives on 'benchmarking' AI ethics

Travis LaCroix[1,2] · **Alexandra Sasha Luccioni**[3,4]

Received: 20 September 2024 / Accepted: 4 March 2025

Abstract
Benchmarks are seen as the cornerstone for measuring technical progress in artificial intelligence (AI) research and have been developed for a variety of tasks ranging from question answering to emotion recognition. An increasingly prominent research area in AI is ethics, which currently has no set of benchmarks nor commonly accepted way for measuring the 'ethicality' of an AI system. In this paper, drawing upon research in moral philosophy and metaethics, we argue that it is impossible to develop such a benchmark. As such, alternative mechanisms are necessary for evaluating whether an AI system is 'ethical'. This is especially pressing in light of the prevalence of applied, industrial AI research. We argue that it makes more sense to talk about 'values' (and 'value alignment') rather than 'ethics' when considering the possible actions of present and future AI systems. We further highlight that, because values are unambiguously relative, focusing on values forces us to consider explicitly *what* the values are and *whose* values they are. Shifting the emphasis from ethics to values therefore gives rise to several new ways of understanding how researchers might advance research programmes for robustly safe or beneficial AI.

Keywords Value alignment · AI ethics · Benchmarking · Unit testing · Metaethics · Moral dilemmas

1 Introduction

Several benchmark datasets have been developed to measure technical progress in artificial intelligence (AI) research, encompassing tasks such as question answering [185], facial recognition [113], machine translation [34], etc. At the same time, the subject of AI ethics—including questions surrounding safety, fairness, accountability, transparency, etc.—has become increasingly prominent as a research direction in the field in recent years. However, there is presently no community-accepted standard for measuring the 'ethicality' of an AI system—i.e., whether the decisions rendered by an AI system are morally 'correct'. That is to say, there is no *benchmark* for measuring whether an AI system is 'ethical' or for comparing the performance (in morally-loaded scenarios) between two distinct models or use cases.

In the paragraphs that follow, we argue that it is, in fact, impossible to develop such a benchmark. Part of the problem arises because the word 'ethics' carries significant philosophical and conceptual baggage. More pressing, members of the AI community are not always cognisant of, nor sensitive to, the subtleties and problems that drive research in moral philosophy. For example, some researchers in AI ethics have suggested that moral *dilemmas*—a type of philosophical thought experiment—may be useful as a verification mechanism for whether a model chooses the ethically-'correct' option in a range of circumstances. But, the use of these dilemmas in the context of benchmarking ethics often fails to maintain sensitivity to, e.g., the purpose of philosophical thought experiments like moral dilemmas [136]. Further problems arise because of the implicit assumptions that AI researchers make about the very nature of ethics—particularly, metaethical assumptions about its objectivity. These insights help clarify why attempts to

✉ Travis LaCroix
travis.lacroix@durham.ac.uk

Alexandra Sasha Luccioni
sasha.luccioni@huggingface.co

[1] Durham University, Durham, UK

[2] Schwartz Reisman Institute for Technology and Society, University of Toronto, Toronto, Canada

[3] Hugging Face (United States), New York, USA

[4] McGill University, Montréal, Canada

Published online: 19 March 2025

benchmark ethics for AI systems presently fail and why they will continue to do so.

Thus, we argue that alternative mechanisms are necessary for evaluating whether an AI system 'is' ethical. These considerations are especially pressing in light of the prevalence of applied industrial AI research. We also propose that it makes more sense to talk about 'values' (and 'value alignment') rather than 'ethics' when considering the possible actions of present and future AI systems. We further highlight that because values are unambiguously relative, focusing on values rather than ethics forces us to consider explicitly *what* and *whose* values they are. This practice has additional downstream benefits for conceptual clarity and transparency in AI research. Therefore, shifting the emphasis from ethics to values gives rise to several new ways of understanding how researchers might move forward with a programme for robustly safe or beneficial AI.

Our method in this paper is couched in the framework of analytic philosophy. As such, we focus on examining the concepts and language employed by researchers when discussing the possibility of benchmarking AI ethics. We bring metaethical considerations to bear on the fact that current approaches to benchmarking ethics for AI implicitly assume a particular metaethical stance—namely moral realism. However, researchers are not typically warranted in doing so; hence, they beg the question about the degree to which the outputs of a particular system are ethical. In light of these theoretical considerations, we highlight the potential social implications—pertaining to, e.g., trust in a deployed system—that are at stake when inadequate benchmarks are used to make unwarranted claims about the (purportedly) ethical features of a particular system. Although our analysis is couched at a theoretical level, we draw attention to practical implications for AI research.

We begin with a discussion of benchmarking in general, highlighting some of the issues recently identified in existing machine learning (ML) datasets and benchmarks (Sect. 2). We then consider benchmarks in the context of ethics for AI systems (Sect. 3) and why they fail. In particular, we discuss a supposed benchmark for ethical AI that has arisen in the context of autonomous vehicles as a particular case study: the 'Moral Machine Experiment' (MME) [155]. We follow with a discussion regarding what values are transmitted via AI research and whose values they are (Sect. 4).

2 Measuring progress in artificial intelligence

Generally speaking, a benchmark can be described as a dataset in combination with a metric—defined by some set of community standards—used for measuring the performance of a particular model on a specific task [184]. Benchmarks are meant to provide a fixed and representative sample for comparing models' performance and tracking progress on a particular task. In this section, we describe some examples of benchmarking results for typical ML tasks and then highlight the myriad ways that have been noted in the literature in which these standard benchmarks give rise to certain issues (2.1). We then discuss how human performance on certain tasks is increasingly used to benchmark model performance and why this approach is illogical given the differences between humans and algorithms (2.2).

2.1 Issues with existing benchmarks

Since its inception, designing tasks and measuring model performance have been central to the field of AI. These continue to be an important part of how members of the AI community compare models. However, despite the ubiquity of benchmarking, major issues have been identified in existing ML datasets and benchmarks.[1] These issues can arise from, e.g., subjective or erroneous labels [87] or a lack of representation, leading to systematic failures across datasets and evaluative approaches [145, 184]. For example, datasets like ImageNet [62] depend on linguistic hierarchies created in the 1980s and include outdated terms such as 'harlot' and 'slattern' [57].[2] At the same time, some of the most commonly-used datasets (including ImageNet) have been shown to contain an average of 3.3% labelling errors [176], with certain classes having error rates up to 98% [151].[3] At best, these issues can affect model performance since they represent noisier data, making it harder for models to learn meaningful representations [186] and for researchers to evaluate model performance properly [176]. Further, this can preserve problematic stereotypes or biases, which are difficult to identify in models deployed in the real world [128, 227]. At worst, they may reinforce, perpetuate, and even generate harms by creating negative feedback loops that further entrench societal structural inequalities [75, 177].

Above and beyond specific datasets, entire AI tasks—such as recognising faces and emotions—have been repeatedly flagged as problematic (often for similar reasons as described above) [44, 205, 206]. Nonetheless, these tasks continue to be used for benchmarking models and

[1] In the context of this paper, we use 'AI' to refer to general approaches in the field pursuing machine intelligence, and we use 'ML' to refer specifically to non-linear statistical approaches within that field.

[2] As another example, the WordNet hierarchy upon which ImageNet is based defines 'queer' as an 'offensive term for a homosexual man', and it links several pejorative semantic relations to this word [160].

[3] If 3.3% sounds like a reasonable error rate, consider that these datasets are often huge. ImageNet contains more than 14 million images, meaning that nearly 1 million of these images—commonly used for training—might be erroneously labelled.

developing entire systems. One such task involves predicting the mortality of different passengers aboard the HMS Titanic [122]. This task has been used for hundreds of tutorials, blog posts, and ultimately published studies [16, 123, 198, 201, 207]. However, whether or not a particular passenger survived is mostly predicted by their gender and the fare they purchased—i.e., their class or social status [40]. So, the task of predicting the fate of passengers on the Titanic is morally dubious—especially when it is done without considering the social inequalities that gave rise to differential mortality rates in the first place.

Consider another example, from the field of computer vision. Oft-used tasks have included applying makeup to images of female faces [47, 117, 144], changing women's clothes from pants to mini-skirts [164, 228], and censoring nude women's bodies by, e.g., covering breasts with a bikini top [169, 200]. Such tasks are ethically problematic because they perpetuate gendered biases and stereotypes, thus reinforcing harmful systems of sexism and misogyny [154]. Even so, these tasks are routinely used as acceptable benchmarks for computer vision models and their results are accepted at leading AI conferences, such as CVPR and ICCV. Although some publication venues—academic conferences and journals—are starting to forward ethical guidelines for both authors and reviewers [21], there is still a general lack of consensus about what constitutes acceptable tasks and applications of ML. This variance exacerbates the fact that it is not obvious that such guidelines will be effective in the first place [139]. Furthermore, creating larger and larger datasets is relatively cheap, but the process of filtering those datasets or 'detoxifying' the models trained on them is expensive [29, 221, 226]. In addition, even when these changes in the direction of 'more ethical' or for a 'common good' are well-intentioned, the lack of conceptual clarity surrounding the targets of such change—i.e., considering what it means to 'be ethical' in the first place—will only compound the issue [56, 93, 167, 208].

In natural language processing (NLP), issues with benchmarks can be more challenging to identify. Still, these may range from unscientific task framing—such as predicting IQ scores based on written text [119] or 'recognising' emotions based on facial expressions [157]—to embedded gender and cultural stereotypes in common NLP benchmarks [33]. For example, in a recent survey of gender biases in NLP models, Stańczak and Augenstein (2021 [204] highlight four key limitations for NLP research[4] (1) gender is often interpreted in a binary fashion, leading to, e.g., misgendering or erasure of non-binary gender identities [19, 76]; (2) NLP research is primarily monolingual, often focusing solely on the English language [54, 129, 219, 220]; (3) biases are typically tested *post hoc*—i.e., after the model has been deployed [162]; and (4) when research explicitly tests for bias (which is infrequent), the evaluation metrics are often incoherent [204]. Thus, even when benchmarks exist for a particular task, researchers lack good baselines for testing ethics considerations in their models—of which bias is one salient example. However, most newly-developed algorithms in this field do not test their models for biases in the first place, and ethical considerations are often ignored.

[4] In this context, biases can be understood as behaviours that involve *systematic discrimination*: against specific individuals or groups (typically in favour of other individuals or groups) [83].

2.2 Benchmarking humans and machines

As mentioned, AI models' performance is increasingly compared to that of humans, with some models reporting 'superhuman performance' on, e.g., game-playing [42, 45, 163, 168, 195, 199, 209], image recognition [107], linguistic tasks [108], etc. However, such comparisons are often misguided (at best) and incoherent (at worst). Recent research has shown that many 'superhuman' language models fail on simple challenge examples requiring compositionality [174], logical reasoning [90], or even simple negation [112]. At the same time, human performance on certain tasks—e.g., diagnoses from X-rays—are often measured by the accuracy of binary outputs (i.e., a particular diagnosis is either positive or negative). In contrast, diagnostic AI models are continuous, including certainty or confidence [88]—this makes it difficult to compare the two, since the decision threshold can change depending on model parameters. Finally, comparing human and machine performance using the same metrics is precarious because metrics such as accuracy, widely used in AI, often fail to correlate with human judgement [32]. Thus, there is a sense in which human performance on tasks is incomparable to computer performance, making any claim of comparison incoherent—not to mention that such comparisons imply 'a narcissistic human tendency to view ourselves as the gold standard' [143].

But given this divergence, it is important to systematically measure progress in AI, either alone or in comparison with 'human-level performance'. However, for this to be possible or meaningful, performance metrics must provide similar conditions for humans and algorithms. An emerging research topic seeks to bridge this gap by establishing more 'equitable' settings for such comparisons—e.g., by imposing constraints such as reduced exposure time for algorithms [84] or a restricted set of label options for humans [69]. For instance, recent work shows that running images through human-like processing filters before feeding them through an algorithm helps 'even the playing field' for both humans and machines [71]. These insights have led to proposals that AI models' performance on standard

benchmarking tasks is not representative of any underlying capacity or lack thereof, given the nature and context of the tasks [78].

Existing proposals have forwarded new evaluation benchmarks that aim at measuring models' robustness and capacity to generalise to new tasks, both from a natural language [39, 51, 229] and a computer vision perspective [111, 170], finding that many models that succeed at existing benchmarks fail at these. Recent work has also proposed alternative approaches such as iterative benchmark development [73] and dynamic benchmarking [126], which endeavour to bring entire fields towards a more nuanced, complex, and informed way of comparing models and measuring progress [66, 196].

So far, we have surveyed some of the practical and theoretical failures of benchmarking in the easy case—i.e., when there is a relatively straightforward answer to questions about model performance, at least in principle. However, even if the issues with existing benchmarks (and their underlying datasets) on well-defined tasks are resolved, these problems severely limit any possibility of benchmarking ethics for AI systems insofar as ethics tasks are rarely, if ever, well-defined. This difficulty is a consequence of the very nature of ethics, as we discuss in the next section.

3 Moral benchmarks for AI systems

As AI systems become increasingly autonomous and more deeply integrated with society, it is obvious that some of the decisions made by these systems will begin to have moral weight. For example, consider a narrow chess-playing algorithm that can only make decisions confined to the action space provided by a chessboard. If the model opens with the Queen's Gambit, this is not a *moral* decision under any definition of morality. In contrast, the decisions made by an autonomous weapon system [9–11, 109, 132, 214], a healthcare robot [5, 6, 55, 197], or an autonomous vehicle [27, 74, 202] may have moral weight. In these cases, the action space may include decision points that we might call 'moral' or 'immoral'—for example, choosing to prioritise one patient over another.

Part of the distinction between a chess-playing algorithm, whose decisions are confined to a particular action space, and an algorithm that acts in the real world is that the decisions made by the latter systems have the potential to impact others. So, in theory, deploying AI systems in the real world logically implies that they will sometimes need to make decisions with moral weight. However, as the action space increases, the set of possible failure modes increases exponentially. Further, the economic promise of AI implies that these systems are increasingly being deployed in society rather than being rigorously tested in the confines of a research lab, thus increasing the risk of harm [138, 150]. Of course, it is not necessary to posit some future science-fiction version of an AI robot acting autonomously in the world to see that the decisions of AI systems may create harm. As a case in point, even narrow AI systems today perpetuate harmful biases, affecting real-world outcomes [7, 50, 213]. And, as mentioned, these decisions may give rise to negative feedback loops, which further entrench those biases (and the harms caused by them) in society [75, 177].

It should come as no surprise, then, that research on ethical behaviour or decision-making in AI systems would attempt to construct a coherent measure for determining whether a system is 'acting ethically'—i.e., whether a decision the model renders is morally 'correct'. Given the historical importance of benchmarks for developing and evaluating AI systems, it makes sense that researchers would try to utilise this tool for evaluating the moral performance of an AI system. However, we argue in this section that benchmarking ethics in this way is impossible. First, we highlight how AI researchers have used moral dilemmas from philosophy as benchmarks for moral performance (3.1) and some recent work criticising this approach (3.2). We then introduce philosophical research in metaethics to show how some substantive claims about the nature of ethics are often taken for granted in discussions of ethical AI (3.3). Finally, we turn our discussion toward real-world distributions to highlight that even if our claims about the nature of ethics turn out to be false, it will still be impossible to benchmark ethical behaviour in an AI system (3.4).

3.1 Moral dilemmas and normative theories

The most common metric for evaluating whether or not a system is ethical is how the algorithm performs on particular moral *dilemmas* [171]. Before we discuss benchmarking ethics using moral dilemmas, we introduce what a moral dilemma is in the first place. To take a concrete example, trolley-style problems are sometimes used to consider certain morally-loaded decisions that autonomous vehicles (AVs) might have to make as these systems become increasingly ubiquitous in society. The trolley problem was originally introduced by Philippa Foot [81]—and later extended by Judith Jarvis Thomson [211, 212]—to consider why it might be permissible to perform some intentional action, A, in situation, S, despite its foreseeable (and undesirable) consequences of A in S.[5] Consider the following scenario.

Bystander at the Switch

[5] This principle dates to at least Aquinas [8]; Foot calls it the *Doctrine of Double Effect* [80]. See also discussion in Kamm [124], Unger [215].

> Suppose there is a trolley heading toward five individuals tied up on the tracks and unable to move. You are near a switch, which would divert the trolley to a separate track, where there is only one individual on the track (also unable to move). You have two (and only two) options:
>
> (1) Do nothing, in which case the trolley is guaranteed to kill the five people on the main track.
> (2) Pull the switch, diverting the trolley onto the side track where it is guaranteed to kill one person.

This standard formulation can be contrasted with the following alternative trolley problem:

> **Bystander on the Footbridge**
>
> Suppose you are on a footbridge above a set of trolley tracks. Below, an out-of-control trolley is approaching five people on the track. The only way to stop the trolley is by dropping something of sufficiently heavy weight onto the tracks to block its path. As it happens, there is a person nearby of sufficiently heavy weight. You have two (and only two) options:
>
> (1) Do nothing, in which case the trolley will kill the five people on the track.
> (2) Push the person off the bridge, thus killing them (but thereby saving the five others).

Each of these is a particular type of philosophical thought experiment, called a moral dilemma [156]. Note that different normative theories from moral philosophy might offer divergent prescriptions (or proscriptions) when these two cases—**Switch** and **Footbridge**—are considered together. In this context, 'normativity' concerns an evaluation or judgement—e.g., that one *ought* to do something. (We will use the phrase 'normative theory' throughout this paper to refer to theories from moral philosophy, without necessarily committing to any claims about 'morality' or 'ethics'.) A 'prescription' can be understood as the provision of a rule to follow or an action to take—i.e., a prescription that one *ought* to ϕ or that one *must*ϕ. In contrast, a 'proscription' is the provision of something forbidden—i.e., a proscription that one *ought not* to ϕ, or that one *must not*ϕ.

Consider a concrete example of how distinct normative theories may offer divergent prescriptions in the same scenario. Certain forms of utilitarianism [24, 159]—a consequentialist normative theory that prescribes utility-maximisation as a reason for action—would recommend acting in *both***Switch***and***Footbridge** because five deaths are obviously worse than one death. On the other hand, a Kantian brand of deontology [125, 130, 131]—a non-consequentialist normative theory which emphasises the importance of duties—would at least say that it is impermissible to act in **Footbridge** since this requires treating a human agent as a means to an end, rather than an end in itself, thus violating the *Categorical Imperative* [125].[6] So, two different normative theories may prescribe (or proscribe) different actions in the same context because they take competing considerations to be important for moral decisions—in this example, consequences on the one hand and duties on the other. Although we have specifically mentioned consequentialism and deontology in this example, the fact that there are many distinct normative theories—both secular and religious—exacerbates the problem that we are highlighting.[7]

In many cases, different normative theories will prescribe the same action (although, possibly for different reasons). However, as we have seen, there may be some tension between the prescriptions of these theories, and moral dilemmas can serve to make these differences salient. Further, moral dilemmas underscore tensions between individual intuitions regarding the rightness or wrongness of an action in a given scenario. In empirical studies, most individuals say they would only act in the case of **Switch**, not in **Footbridge** [38, 173]. Thus, both the prescriptions of normative theories *and* common intuitions about the permissibility of an act may vary.[8] The point is that a moral dilemma is a tool for philosophical analysis used to bring these tensions to the fore.

Note that the key point here is not to critique the normative theories themselves. The use of moral dilemmas in philosophical thought experiments is precisely that these dilemmas are knife-edge cases which can elicit intuitions about the potential shortcomings of how a particular normative theory approaches the dilemma. For example, one might find that consequentialism's emphasis on outcomes or deontology's adherence to universal rules leads to situations that most people would call unethical in certain circumstances.

[6] One formulation of the categorical imperative states it is never permissible to use a human agent as a means to an end. It is less obvious whether this imperative would also proscribe acting in **Switch** However, Thomson [212] argues that there is a sense in which **Switch** still uses a human agent as a means to an end and thus would be impermissible by Kantian deontology.

[7] Some authors have rightfully pointed out the monopoly of Western ethics for setting norms in emerging AI technologies; see, e.g., Elmahjub [70].

[8] Of course, how people respond to abstract philosophical dilemmas on questionnaires may be quite different from how they act in the real world [37, 173].

Part of the purpose of a moral dilemma (as a type of philosophical thought experiment) is to focus attention on the morally-*salient* features of the dilemma [41, 63–65] without getting bogged down by the pre-theoretic baggage that individuals may carry. In the case of the trolley problem, Foot's [81] original target of analysis is abortion (not trolleys). However, the thought experiment is useful precisely because of the supposed tension (at least in western analytic philosophy) between *emotion* and *rationality* [82, 115, 116, 203]: people are less likely to carry pre-theoretic baggage about trolleys than they are about abortions. Therefore, the thought experiment gets to the core of a moral issue in applied ethics while abstracting away from the actual (morally-loaded) target [136].

Hence, rather than critiquing one or another particular normative theory, our point is to map the use of moral dilemmas to see whether this tool can be aptly applied to AI ethics in the context of benchmarking—i.e., determining whether or how ethical the outputs of an AI system are. Despite the conceptual purpose of dilemmas in moral philosophy—i.e., as thought experiments or 'intuition pumps' [63–65]—AI researchers have begun to use these dilemmas as validation proxies for whether a model is ethical. In the remainder of this section, we discuss why this is a mistake.

3.2 Moral machines

As we have seen (Sect. 2), what we might call the 'standard model' for measuring 'progress' in AI research involves benchmarking. Thus, it stands to reason that to determine whether (1) a choice made by a particular model in a morally-loaded scenario is the (morally) 'correct' one, (2) one model is 'more' moral than another, or (3) a model is increasingly 'moral' when subjected to further training, it appears that researchers need a *benchmark* for measuring the 'ethicality' of a model. Logically, then, for such a task to be successful, we would require an ethics dataset—either general-purpose or task-specific—and a metric for measuring model performance relative to that dataset.[9]

To take a specific example, trolley-style moral dilemmas, like **Switch** and **Footbridge**, have been widely discussed in machine ethics and AI research in the context of possible (low-probability but high-stakes) situations in which an autonomous vehicle (AV) may be placed.[10] Suppose that a fully-autonomous vehicle must 'choose' between killing five pedestrians or swerving into a barrier, killing the driver in the process. Functionally, this scenario is equivalent to a trolley problem, in that an actor must make a choice, the consequences of which will involve one death or several.

Perhaps the most well-known instantiation of this dilemma in an AI context is the *Moral Machine Experiment* [155] (MME): a multilingual online 'game' for gathering *human* perspectives on (hypothetical) moral decisions made by a machine intelligence. Participants are shown several unavoidable accident scenarios with binary outcomes and are prompted to choose which outcome they think is more *acceptable*. These include 'sparing humans (versus pets), staying on course (versus swerving), sparing passengers (versus pedestrians), sparing more lives (versus fewer lives), sparing men (versus women), sparing the young (versus the elderly), sparing pedestrians who cross legally (versus jaywalking), sparing the fit (versus the less fit), and sparing those with higher social status (versus lower social status)' [13, p. 60]. The MME appears to provide a type of benchmark in the following sense: the *dataset* is the set of data collected online from humans in response to the hypothetical scenarios posed; the *metric*, then, could be how closely the decision of a model accords with the data for a given scenario—i.e., human responses *on average*.

However, this approach to the problem of creating 'moral' AI systems is highly pernicious. First, the MME data are *descriptive* rather than *normative*. That is, the data do not tell us (or a model) anything about how one *ought* to act in a given scenario; instead, the data offer a *description* of how people (hypothetically and on average) would (or say they would) *want* an autonomous vehicle to act in such a scenario. As a result, using these data for benchmarking a new algorithm is a type of *fallacy*—i.e., the logical error of deriving an 'ought' from an 'is' [114, 166]. The error of reasoning arises from the implication that since people say they would *want* an AV to act in this way (a descriptive claim), it follows that the machine *ought* to act in this way (a normative claim).[11]

Second, the thing being measured against the MME data is not whether a decision is, *in fact*, ethical, but how well a decision *corresponds to the opinions of a particular set of humans, on average*. For an ethics benchmark to be useful, it must provide data for the *de facto* morally-'correct' decision in a given scenario. The MME data provide a mere *proxy* for

[9] Recall that, in Sect. 2, following Raji et al. [184], we described a benchmark as a dataset in combination with a metric.

[10] See, for example, Allen et al. [4], Wallach and Allen [218], Pereira and Saptawijaya [180, 181], Berreby et al. [26], Danielson [59], Lin [146], Malle et al. [153], Saptawijaya and Pereira [192, 193], Bentzen [25], Bhargava and Kim [27], Casey [46], Cointe et al. [52], Greene [95], Lindner et al. [147], Santoni de Sio [191], Welsh [222], Wintersberger et al. [223], Bjørgen et al. [30], Grinbaum [96], Misselhorn [161], Pardo [178], Sommaggio and Marchiori [202], Baum et al. [18], Cunneen et al. [58], Krylov et al. [134], Sans and Casacuberta [190], Wright [224], Agrawal et al. [2], Awad et al. [12], Banks [15], Bauer [17], Etienne [72], Gordon [92], Harris [106], Lindner et al. [148], Nallur [171].

[11] Although Awad et al. [13] are careful to highlight that the MME is only supposed to be descriptive, several authors of the MME use these data to propose a 'voting-based system for ethical decision making' [175], which is clearly a normative project [136].

this target: namely, a sociological fact about how some set of human agents annotates a particular set of decision problems, on average. Such proxies are especially harmful when the researchers who use them do not maintain sensitivity to the differences between the proxy and the target. This is, in effect, a value-alignment problem [135, 137], which we will discuss in more detail in Sect. 4.

Third, although there are intrinsic reasons why we might want AI systems to be capable of acting ethically, the AV case brings to light a different type of value-alignment problem. Namely, for-profit corporations have some market incentives for designing 'ethical' AI since humans (i.e., consumers) will likely be more trusting of an autonomous agent (i.e., a product) if it is known to possess a set of moral principles intended to constrain and guide its behaviour [35]. However, suppose that the (in fact) 'ethical' decision between killing five pedestrians and swerving into a barrier, thus killing the passenger of the AV, is to swerve. Human consumers may be less willing to purchase a product that may choose to kill them, even if it is the 'most ethical' decision. Indeed, a human consumer may be more willing to purchase a product that follows the *pseudo*-moral imperative: always prioritise the passenger's wellbeing. Therefore, the companies that design these models have perverse profit-maximising incentives when designing 'ethical' AI. We will discuss this in more detail in Sect. 4.3.

The MME exemplifies a trend that attempts to use moral dilemmas from philosophy as benchmarks for ethical AI. For example, Nallur [171] suggests that if some model implementation can 'resolve a dilemma in a particular manner, then it is deemed to be a successful implementation of ethics in the robot/software agent' (p. 2382). Additionally, Bjørgen et al. [30] argue that certain types of ethical dilemmas—including the trolley-style problems discussed above—'can be used as benchmarks for estimating the ethical performance of an autonomous system' (p. 23). Similarly, Bonnemains et al. [36] argue that 'it seems legitimate to use some [moral dilemmas] as a starting point for designing an automated ethical judgement on decisions' (p. 43) because classic moral dilemmas have already been used as a basis for ethical reasoning. And, this reasoning extends well beyond the particular use of trolley-style problems for reasoning about ethical decision-making in autonomous vehicles; for example, Lourie et al. [149] introduce a dataset of ethical dilemmas, which they suggest 'enables models to learn basic ethical understanding'. However, LaCroix [136] argues that using moral dilemmas for benchmarking involves a category mistake. *Moral dilemmas have no right answer, by design*.

To make the point concrete, suppose an autonomous vehicle is trained on and benchmarked against MME data. Suppose that the system scores highly on this benchmark, and the company advertises the system as being more ethical than its competitor. Nothing about this situation implies that the system is ethical, because the benchmark depends upon a moral dilemma. What is being measured is accordance with the data, so the claim toward 'being ethical' depends inherently upon the degree to which we think that the survey data align with some matter of fact about ethics—i.e., how good of a proxy those data are for the moral matter of fact.

However, this presupposes a metaethical stance—i.e., that there *are* matters of fact about ethics. Regardless of whether moral realism is true or false, what our analysis is attempting to clarify is a lack of care to subtleties surrounding metaethics. Moreover, neglecting these considerations can have significant social implications when we reflect upon the incentive structures of those entities that create and deploy AI systems. Namely, if the benchmark becomes the target, but the target is inapt, then systems will be touted as more or less ethical despite that ethics is not being measured by the benchmark. This situation is tantamount to misinformation (or hype) about the functioning of these systems; however, in contrast to other use-cases for benchmarking—e.g., standardised tests, labelled images, etc.—the *normative* context carries with it significantly higher stakes when it comes to getting it wrong. To wit: hand-annotated labels are a poor proxy for describing the complexity of images; standardised tests are a poor proxy for describing the complexity of intelligence; so too, human intuitions about moral dilemmas are a poor proxy for ethical behaviour. Moreover, the ethics case is unlike other benchmarking instances insofar as it is not clear—and, therefore, we cannot simply take for granted—that there is any objective truth about whether something 'is' ethical.

Thus the question that researchers take themselves to address is how to determine *whether* the decision chosen by the system is 'in fact' moral. From the perspective of AI research, it appears that this problem is merely a matter of choosing a metric by which performance on the system can be measured and then determining whether or not the algorithm in question is successful on *that* metric. Once the metric is determined, standard benchmarking techniques may apply such that one algorithm performs better than (or, 'is more ethical than') another. The question then arises how we are supposed to *know* whether the decision chosen by the system is 'in fact' moral—i.e., *how* ethical are the decisions made by the algorithm? We argue that this question is question-begging (at best) and incoherent (at worst) by appealing to philosophical research in metaethics. We first give a very brief overview of some distinct metaethical theories before applying the insights of these theories to the case of benchmarking AI ethics.[12]

[12] The brevity of this overview entails that we have focused narrowly on western philosophical theories. It is worth noting that diverse global

3.3 Ground truths for moral benchmarks

Metaethics is the branch of moral philosophy that seeks to explain the very *nature* of ethics.[13] Moral realism is a meta-ethical view which holds that moral properties exist [194]. A realist about ethics would hold that moral claims purport to report *facts*—i.e., about the world—and are true when they get those facts correct. For example, if I say 'murder is wrong', I am making a normative claim. A moral realist would hold that this proposition is either true or false, regardless of, e.g., social norms or conventions. And, whether this proposition is true or false depends upon some matters of fact—i.e., about the world—independent of me and my views.[14] For benchmarking to make sense in the first place, there must be some ground truth against which one can compare the outputs of one's model. If there were not, then ethics benchmarks would not be measuring ethical performance; instead, the would be measuring, e.g., accordance with described preferences of individuals, on average. Thus, by assuming that ethics is the sort of thing that can be benchmarked, researchers tacitly assume that there is a ground truth—i.e., that there are moral facts, which can be true or false, and that we have epistemic access to those facts. This 'commonsense' view of morality presupposes the existence of objective values.

However, this point is not to be taken for granted. It is highly contentious whether there is any such ground truth in ethics, even amongst experts in the field. For example, *non-cognitivists* about ethics think that moral claims do not express propositions; thus, such claims are not *truth-apt*—i.e., similar to an exclamation or a question, moral claims are not capable of being true or false. One particular brand of non-cognitivism—'emotivism'—likens moral claims to an emotional expression of one's attitude toward some action or set of actions [14]. On this view, an ideal moral benchmark could only measure accordance with individuals' attitudes toward an action based on their emotional expression, rather than some objective fact about ethics simplciter.

Another prominent form of anti-realism about ethics is *error theory*, which holds that all moral claims are *false* (because there are no objective moral values) [152]. Positing the existence of objective facts about ethics may turn out to be nothing more than 'a useful fiction' [121], 'an error' [152], 'a collective illusion' [188], a 'function of social conventions' [103–105], or simply a 'network of attitudes' that is projected onto the world [31]. Each of these metaethical theories brings with it a *distinct* standard for benchmarking. Importantly, however, none of these views would entail (or in some cases allow) that what is being measured is accordance with some objective principle, as is presupposed by a moral realist.

If it turns out that moral realism is false, then benchmarking *ethics* would be impossible because there is no matter of fact—i.e., no ground truth—about ethical claims against which one can benchmark a model. There may be facts of the matter regarding, e.g., social conventions; however, researchers who attempt to benchmark ethics do not appear to have this in mind. Rather, the target seems to be *moral facts*—which, it turns out, may not exist. The point here is not whether moral realism is true or false. The point, instead, is that 'moral realism is true' is a substantive (and contested) claim that cannot be taken for granted. However, this is precisely what is taken for granted when researchers assume that they can benchmark the ethicality of a decision made by their model.

As with the benchmarking issues discussed in Sect. 2, the real problem with benchmarking ethics concerns taking substantive claims for granted and unreflectively applying vague concepts to a problem with potentially significant real-world consequences. Moreover, when considering the downstream effects of work that claims to benchmark ethics, there seems to be a salient distinction to make between a company suggesting that a model performs well on language tasks (and so 'understands' language) or on standardised test tasks (and so 'has' a certain IQ) versus a model that performs well on an ethics benchmark (and so 'is' ethical). The potential social risks that arise from misinformation of the latter type appear more harmful than the former by garnering unearned trust in the functioning of the system. Hence, in addition to inheriting all the standard problems of benchmarking in general, purported benchmarks of ethical behaviour suffer from additional, and potentially insurmountable issues.

Note that there is a sense in which ground truths do not require metaphysical truth: one might claim that although unicorns do not exist, it is still possible to create a machine vision system to correctly distinguish between, e.g., horses, zebras, and unicorns. Should not the same be true about morality? However, this is a faulty analogy. In the unicorn case, the 'ground truth data' correspond to, e.g., metaphysically-real illustrations of the metaphysically-unreal creature. This is unlike the claim that moral anti-realism could still allow benchmarking a model on how accurately its

and comparative ethical perspectives (both secular and religious) may offer different worldviews on metaethical and normative levels. Some recent work in the philosophy and ethics of AI has begun to engage more robustly with these perspectives; see, e.g., Goltz et al. [91], Elmahjub [70], Chen and He [49]. That said, our main claim still stands in this wider context: in order to benchmark the morality of AI systems, one presupposes a meta-ethical stance—i.e., that such-and-such normative theory is correct.

[13] Unlike normative ethics, which asks questions like 'what ought I to do', metaethics is primarily concerned with questions surrounding ethical concepts—e.g., what does a normative word like 'ought' mean?

[14] At least according to certain theories of truth. See [89].

classifications agree with social conventions; however, this is precisely the point. Researchers who seek to benchmark ethics are (wittingly or unwittingly) measuring a proxy rather than the true target. It is important to understand when such substitutions take place to avoid overreaching claims about whether or how ethical one's model is. 'Social conventions' lack the supposed objectivity of 'ethics'. So, although it would perhaps be possible to benchmark social conventions, these are highly relative. Awareness of this fact underscores the need for diversity in AI research because homogeneous groups of researchers are liable to take into account only those social conventions that are salient to them. Hence, a lack of diversity can (and does) lead to homogeneous sets of values.

Further, it matters less what the concept is, and more how accurate the data are. One can have accurate data on fictional objects or entities precisely when there is relative agreement about those entities—e.g., what counts as a unicorn. In contrast, if there are no moral facts, one cannot have accurate data on morality. Again, at best, you might have data about *preferences*, *values*, *conventions*, etc. However, none of these concepts retain the veneer of ontological validity that, we contend, is presupposed in discussion about ethics in the context of AI.

In addition, even if moral realism turns out to be true, thus vindicating the assumptions made by some members of the AI community, benchmarking ethics will still be impossible with current approaches because of the disconnect between the distribution of examples that models see in training and the distribution of states of the real world. Namely, the *long tail problem*.

3.4 A long tail problem

The long tail problem is a longstanding issue in the field of AI. In effect, there are a potentially infinite number of states an AI system might face in the real world, and it is impossible to represent every contingency in the training data. Although gathering data about common objects, contexts, or situations is relatively easy, doing so for uncommon ones is difficult precisely because of their rarity. However, 'rare' does not mean 'impossible'. Following the theory that 'what-ever can happen will happen if we make trials enough' [61], as models are deployed in the real world, it becomes increasingly plausible that they will encounter objects and situations on which they were not trained. Namely, any event with non-zero probability is an actuality in the limit. Even applied AI techniques, like adversarial generation—i.e., training a separate model to artificially generate training data that does not exist in the real world [230]—will not solve this problem because it is impossible to account for all potential scenarios and situations. In practice, these data generation techniques are often coupled with user-defined heuristics, such as compelling a model to abstain from proposing a classification if its confidence threshold is too low or simply removing problematic categories. For example, when Google's AI-based photo-tagging feature labelled two African Americans as 'Gorillas', they removed that particular category from the options available to the model [216]. Nonetheless, both of these approaches are brittle and fail to generalise for the multitude of real-world situations and problems that AI systems face.

Thus, even if we ignore the fact that benchmarking ethics requires significant presuppositions about the nature of ethics (which AI researchers are not warranted to make), the long-tail problem makes benchmarking ethics impossible, regardless of whether there is a ground truth against which a model might be benchmarked. Part of this is the distinction between actions spaces containing decisions with or without moral weight. To go back to our original example, if a chess-playing algorithm has not seen some set of moves, and responds sub-optimally, the worst possible thing that can happen is that the algorithm loses a game of chess. Although this outcome may not be ideal for the researchers who trained the model, it has little real-world consequence. In contrast, when a model encounters a situation that it has not seen before, and its action space includes acts that we would call 'immoral', this can have real-world consequences. Therefore, low-probability but high-risk events pose unique challenges in ethical contexts. This problem is difficult even when there is an objectively correct answer, but as we have seen, some (possibly all) morally-loaded situations have no such claim to objectivity. Thus, the long-tail problem prevents the coherence of benchmarking in the context of ethics even in the possible world in which ethics has some ground truth.

The conceptual difficulties surrounding the very nature of ethics are further exacerbated when researchers are not attentive to them. Although the objectivity of ethics is contested, we suggest that values are unambiguously relative. Therefore, in the next section we suggest that *values*, rather than ethics, are a more appropriate target for research on safe and beneficial AI.

4 The values of AI research(ers)

Given the increasing influence of AI systems on the world around it and the impossibility of benchmarking ethics, it is necessary to investigate the tacit (often value-laden) aspects of model creation and deployment. Considering the values embedded in models is especially important because these can have major downstream impacts on the products and applications in which they are integrated, despite not

being explicitly defined or communicated. In this section, we briefly summarise a description of the value alignment problem, which gives rise to two key questions: *What* values are encoded in AI research? And, *whose* values are they? We suggest that shifting conceptual focus from *ethics* to *value* forces us to engage with these questions in a more robust way than benchmarking ethics allows because mitigating a misalignment of (some set of) values cannot take for granted what or whose values are under consideration. Hence, we trade ambiguity for specificity.

4.1 The structure of value alignment

Instead of attempting to benchmark how ethical and AI system is, we propose that it is more fruitful to consider the degree to which an AI system is aligned with (some set of) values. The value alignment problem is standardly described as the problem of ensuring that AI systems (outputs, behaviours, decisions, goals, objectives, etc.) are aligned with the values (objectives, intentions, etc.) of humanity [50]. Gabriel [85] differentiates between the *normative* component of the problem—what values ought to be encoded in an AI system?—and the *technical* component of the problem—how do we encode those values?—However, this standard description is too vague to be useful [137].

A more practicable description of the value alignment problem for artificial intelligence is given by the following *structural* description:[15]

> The value-alignment problem is a problem that arises from the dynamics of multi-agent interactions involving the delegation of tasks from one actor (a human principal) to another (an AI agent).[16]
>
> This problem can arise whenever
>
> (a) The agent's objective function is misaligned with the true objective of the principal(s); *or*,
> (b) There are informational asymmetries between the principal and the agent.

According to the structural definition, there are three key axes along which value misalignment may arise: misspecified objectives, informational asymmetries, and relative principals.

As we have argued, a benchmarking approach to AI ethics presupposes an objective standard (i.e., a meta-ethical stance) against which and AI system can be benchmarked in the first place: what is *meant* by 'ethical' is taken for granted. In contrast, when we consider 'values' and 'value alignment' no such presuppositions are made because values are inherently relative—a fact that is reflected in the structural definition of value alignment. Hence, in order to mitigate potential misalignment, it is necessary to make explicit *what* are under consideration, and *whose* values they are.

4.2 What values are encoded in AI research?

Models and algorithms carry values encoded by the researchers and institutions that created them. However, these values are often not clearly stated during the peer-review process or subsequently, once the research is formally published. In a recent study, Birhane et al. [28] analysed 100 highly-cited ML papers to identify their intrinsic values. They found that the most common values underlying this research include generalisation, efficiency, interpretability, and novelty—although, these are rarely made explicit. Here, we examine two of the most prevalent values identified in the study: *performance* and *building upon prior work*. We discuss their repercussions on the field's priorities as a whole and the power dynamics that drive them.

Birhane et al. [28] report that the most common value held by the ML research community—present in 87% of the papers analysed—is *performance*. However, benchmarks are the main mechanism for tracking and reporting performance improvements, and we have already seen (Sect. 2) that benchmarks have significant and well-known issues. Another known issue with this performance-centric value is that training higher-*performing* models often entails training *larger* models, given current paradigms in deep learning. However, requirements of size make performance contingent on access to ever-increasing quantities of data and computing power, which is increasingly unsustainable from an economic, technical, and environmental point of view [20, 210].[17] A purely performance-focused mindset also adversely affects researchers from countries and

[15] This structural framing, although somewhat non-standard, is gaining prominence in the recent literature which highlights an analogy between value alignment and the principal-agent problem in economics [98, 137], which increasingly focuses on the structural features of value alignment as a multi-agent problem [79]. This approach informs recent discussions of cooperative inverse reinforcement learning [100], incomplete contracting [99], and game-theoretic approaches that leverage informational asymmetries as a tool for alignment [86, 97].

[16] Note that the use of "agent" here follows the language of the principal-agent framework from economics, which is distinct from other uses of "agent" in philosophy—i.e., referring to the capacity for autonomous action—or the recent literature on AI agents—referring to an autonomous system that interacts with its environment. In this context, the term "agent" refers to an entity that acts on behalf of another party (the principal) by performing a task, making a decision, etc. (regardless of whether this is done autonomously).

[17] For example, Thompson et al. [210] estimate that it would take an additional 10^5 times more computation to achieve an error rate of 5% for ImageNet, based on the current trend of computing requirements for ML. (The present error rate was estimated at 11.5%.) This increase

regions with no access to large-scale computing infrastructures or expensive hardware. This disproportionate disadvantage further amplifies the extant power dynamics within the field [165]. Finally, since performance is so highly-valued in the research community, this creates a negative feedback loop: undue emphasis on performance measures sways the course of subsequent research and influences the directions pursued by others, thus further orienting the field in the direction of pursuing performance as opposed to other, more varied pursuits [68]. There are currently limited mechanisms for flattening the exponential need for compute resources. And, the efficiency of models is not taken into account during their benchmarking.[18] Although alternative approaches are possible—for example, methods for improving neural networks' efficiency [48, 225] and developing more optimised hardware accelerators [182]—these are not currently mainstream endeavours.

The second most prevalent value identified value by Birhane et al. [28] is *building on past work*, which often is (explicitly or implicitly) bound up with valuing *novelty*. Indeed, the structure adopted by many ML papers hinges upon discussing similarities or differences to related works without questioning or critiquing them [140]. The same consideration applies to datasets and benchmarks, which persist despite their shortcomings (including lack of applicability to any real-world deployment of the proposed algorithms) [184]. Even in cases where societal impacts are meant to be mentioned—such as the increasingly-common 'broader impact' statements or checklists now appearing in conference submissions—these statements often fail to address negative societal consequences, keeping any remarks high-level, abstract, or vague [172]. These difficulties have also contributed to a 'reproducibility crisis' in the field: endeavours that aim to reproduce ML research have systematically found that many peer-reviewed papers are missing information necessary for reproducibility [67]. Sometimes these omissions are minor, such as failing to report random seeds and hyperparameter values; however, they can also be significant—e.g., not sharing data and code [110, 183]. However, if past research is impossible to reproduce, it will also be impossible to build upon it (unless past results are taken for granted). Thus, even supposedly marginal details, like random seeds, can have significant downstream effects on future work since the results of past work may be entirely contingent upon these details.

The two values described above are especially pervasive in the field of large language models (LLMs), whose size has drastically increased in recent years: recent models boast progressively more parameters, which are now in the trillions [77]. However, descriptions of these models traditionally emphasise (1) their performance on the same set of benchmarks and (2) that their parameter-count is bigger than that of previous models. Certain relevant aspects of the model—e.g., training time, energy consumption, or compute costs—are often ignored.[19] This lack of transparency regarding the negative impacts of ML models, with an emphasis on those deemed positive by the community at large (e.g., performance, novelty, etc.), further entrenches the presumed contributions of ML while sweeping the cost of these contributions under the rug. Furthermore, when researchers do criticise these models' shortcomings, they may be penalised by the very institutions whose business models hinge upon their success [60]. All this is to say that the values that are encoded by AI research are inherently relative, so it is crucial to consider *whose* values models encode.

4.3 Whose values are encoded in AI research?

In the history of AI research, the computational constraints of the late 1980s and early 90 s forced researchers to make primarily *theoretical* progress on toy datasets or mathematical analysis [22, 142, 187]. This focus shifted in the early 2010s when it became possible to train a deep neural network on a fairly large dataset using a single graphics processing unit (GPU) server [133]. This breakthrough marked a new era in AI when it was possible for researchers to train models on local machines while making progress on datasets such as ImageNet [62] and MNIST [141]. This era did not last, however. In the last decade, the computing needs of AI have grown significantly, and most deep neural networks need to be trained on multiple GPUs, now measured in the hundreds or thousands [179].

This resource-intensive focus has contributed to a major shift in the power dynamics of the field insofar as it puts for-profit technological companies with large amounts of compute at an advantage compared to smaller companies and academic institutions [127]. For example, Birhane et al. [28] found that 79% of the highly-cited papers they analysed were written by authors with ties to corporations. This figure is corroborated by previous work that has analysed the increased presence and power that big tech companies wield in the field of AI [1, 3]. Given the increased contributions of

would produce an additional 10,000 pounds of carbon emissions and cost millions of US dollars.

[18] For example, a model that achieves an increased accuracy of 0.5% on ImageNet while requiring one month of compute is still considered 'better' than a model achieving an increase of 0.45% with only one week of compute.

[19] For instance, while the paper accompanying GPT-3—a recent LLM with 175 billion parameters—reported its performance extensively on 42 'accuracy-dominated benchmarks', the authors provided no details on training time or compute costs [43].

for-profit companies to AI research, it is important to keep track of their effect on research directions in the field. This situation constitutes a sort of value-alignment problem—namely, the problem of aligning the 'goals' of AI systems with human values [50, 85, 189]—insofar as the incentives and goals of corporations may not align with a common good or the values of humanity, writ large [138]. However, tracking these effects is difficult given the current lack of transparency around values driving industrial AI research.

Concretely, the influence of for-profit corporations on AI research can vary, ranging from the seemingly harmless funding of academic research (provided that it aligns with a company's interests) to employing teams of researchers dedicated to pursuing in-house research. In the latter case, confidentiality may be protected by non-disclosure agreements, intellectual property laws, and multiple levels of compliance. Since salaries paid by academia and industry are increasingly disparate, more and more talented students and researchers are leaving academia for the prosperity promised by industry research, further widening the gap between the two camps [158]. Abdalla and Abdalla [1] highlight that the strategies employed by large technological corporations to maintain their freedom to develop and deploy AI tools and products while avoiding accountability and increased legislation are comparable to those employed by Big tobacco for decades to downplay the harmful effects of cigarettes. These techniques range from maintaining a socially acceptable image to influencing government legislation [1]. These tactics are made possible by the extensive financial resources companies have, which far surpass the funding of academic institutions.

In the realm of moving research in an 'ethical' direction, ethics *guidelines* have proliferated in recent years [118]. These guidelines, codes, and principles come from various sources, including for-profit corporations. And, it has been pointed out that this implies that these stakeholders have a vested interest in shaping policies on AI ethics to fit their own priorities [23, 94, 118, 139, 217]. Our proposal to shift focus from ethics to values is differentiated from the normative principles approach of ethics guidelines insofar as such rule-based approaches are top-down and still presuppose a meta-ethical stance, insofar as such principles attempt to *codify* individual values into moral standards. Considering 'values' and whether or the degree to which a system is 'value aligned', on our view, avoids some of the theoretical issues arising in the context of ethics guidelines insofar as considering the values encoded in AI research and the values of individuals and communities with which those systems are supposed to align) does not entail a standard against which the system purports to be benchmarked.

In the context of applied ethics, the current emphasis in AI research has been on the *technical component* of problems such as value alignment; this has the unfortunate consequence of ignoring the difficult work of determining *which* values are appropriate in the first place—i.e., the *normative component* of value alignment [85]. Furthermore, moral and political theory are deeply interconnected with the technical side of the AI alignment problem. And, as we argued in Sect. 3, second-order ethical commitments are often taken for granted by AI researchers. More difficult still, suppose we discovered or determined that, e.g., *utilitarianism* is the objectively-correct normative theory. Even then, the utility considerations upon which this theory depends will always be relative to some frame. The theory prescribes maximising utility, but we must still ask: utility *for whom?* And, it is important to understand that *no* decisions made by researchers are value-free; this work is never neutral. As Green [93] emphasises, '[b]road cultural conceptions of science as neutral entrench the perspectives of dominant social groups, who are the only ones entitled to legitimate claims of neutrality'.[20]

When researchers say that such-and-such model 'is' ethical, or they unreflectively deploy normative terms like 'social good', this leaves certain metaethical and normative presuppositions and commitments implicit. Engaging in a discussion of *values* rather than ethics brings these commitments to the fore. Researchers are not warranted to say that *any* model is ethical unless they explicitly define what they mean by 'ethical'—high performance on a nonsense benchmark will not suffice. And, even then, the definition will be subject to criticism (if the history of Western philosophy is any indication).

5 Conclusion

AI is still a relatively new and rapidly changing field, and we have already seen some movement toward more socially-minded research and practice in recent years. However, we can still improve efforts to increase transparency and accountability within our community. For instance, AI researchers need to be mindful and reflective regarding the capabilities and limitations of both models and benchmarks by, e.g., reporting metrics other than accuracy and carrying out more in-depth error analysis can paint a more nuanced picture of performance, highlighting what models have yet to succeed on and sharing failure cases alongside capabilities. Also, while checklists covering topics ranging from copyright to the broader impacts of AI systems are starting a mandatory part of the submission process for many ML conferences (e.g. NeurIPS and ICML), continuing work to make these checklists more thorough (for instance, including

[20] See also, [53, 101, 102, 120].

question regarding values and biases) can help make these issues more transparent and allow us to get a better idea of the value systems present in the AI community.

Benchmarking ethics would require a ground truth about ethical claims. A 'commonsense' view of morality presupposes that ethics is objective. Researchers in AI have taken this view for granted. In this work, drawing upon research in moral philosophy—including normative ethics and metaethics—we have provided a positive argument in favour of shifting the discourse of AI ethics toward talk of values.

Broadly construed, values are basic and fundamental beliefs that guide action. They have normative weight, insofar as an individual ought to do things that align with their values. Since values are shaped by cultural, social, and personal experiences, they can vary significantly across different societies and adapt as cultural norms evolve. In contrast to structured normative frameworks—e.g., consequentialism, deontology, virtue ethics, etc.—which are supposed to hold universally, values are inherently flexible and context-dependent. Moral theories can struggle to accommodate complex, real-world scenarios that require balancing competing values; hence, as AI systems become more pervasive in everyday life, we should expect them to fail to provide appropriate guidance for particular situations. This fact is precisely what is suggested by the discussion of moral dilemmas above, further highlighting that these are inappropriate tools for benchmarking moral behaviour.

We suggested above that talk of ethics provides a misleading air of objectivity and universality—features that cannot be taken for granted when considering metaethical perspectives on normative theories. Hence, although the average individual may be misled by claims that such-and-such system 'is' ethical, describing the systems behaviour in the context of value alignment forces one to address additional questions—most notably, with what values does the system align, and whose values are these.

By highlighting the distinction between supposedly-objective concepts, such as commonsense views about ethics, as opposed to relative ones, such as values, we have proposed that shifting the discourse would contribute towards making our field more transparent. Although, on its face, this may seem like a modest proposal, it would have significant consequences for how discussions of ethical AI proceed and provide further opportunity for positive change in our field.

Acknowledgments TL would like to thank Elinor Bell-Clark, Fintan Mallory, Greg Lusk, Brendan Kelters, and Alex Campolo for helpful discussion. Some of the research of this paper at the American Philosophical Association, Pacific Division meeting (Vancouver, 2022) and the Philosophy Department Colloquium at Dalhousie University (Halifax, 2023); many thanks to the audiences at these venues for their engagement and to Duncan Purves for commentary.

References

1. Abdalla, M., Abdalla, M.: The grey hoodie project: big tobacco, big tech, and the threat on academic integrity. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp. 287–297. AAAI/ACM, New York (2021)
2. Agrawal, M., Peterson, J.C., Griffiths, T.L.: Scaling up psychology via scientific regret minimization. Proc. Natl. Acad. Sci. USA **117**(16), 8825–8835 (2020)
3. Ahmed, N., Wahed, M: The de-democratization of AI: deep learning and the compute divide in artificial intelligence research, pp 1–52. arXiv pre-print arXiv:2010.15581 (2020)
4. Allen, C., Wallach, W., Smit, I.: Why machine ethics? In: Anderson, M., Anderson, S.L. (Eds.) Machine Ethics, pp. 51–61. Cambridge University Press, Cambridge (2011)
5. Anderson, M., Anderson, S.L.: Ethical healthcare agents. In: Sordo, M., Vaidya, S., Jain, L.C. (eds.) Advances Computational Intelligence Paradigms in Healthcare 3. Studies in Computational Intelligence, vol. 107, pp. 233–257. Springer, Berlin (2008)
6. Anderson, M., Anderson, SL., Armen, C.: Medethex: a prototype medical ethics advisor. In: Proceedings of the Eighteenth Conference on Innovative Applications of Artificial Intelligence (IAAI-06), pp. 1759–1765. AAAI, Boston (2006)
7. Angwin, J., Larson, J., Mattu, S., Kirchner, L.: Machine bias. ProPublica, May 23: np (2016). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
8. Aquinas, T.: Summa theologiæ (1485)
9. Arkin, R.C.: Governing lethal behavior: embedding ethics in a hybrid deliberative/reactive robot architecture—part I: motivation and philosophy. In: HRI'08: Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction, pp. 121–128. Association for Computing Machinery, New York (2008)
10. Arkin, R.C.: Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture—part II: formalization for ethical control. In: Proceedings of the 2008 Conference on Artificial General Intelligence 2008: Proceedings of the First AGI Conference, pp. 51–62, Amsterdam. IOS Press (2008)
11. Asaro, P.: Autonomous weapons and the ethics of artificial intelligence. In: Liao, S.M. (ed.) Ethics of Artificial Intelligence, pp. 212–236. Oxford University Press, Oxford (2020)
12. Awad, E., Dsouza, S., Bonnefon, J.-F., Shariff, A., Rahwan, I.: Crowdsourcing moral machines. Commun. ACM **63**(3), 48–55 (2020)
13. Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., Rahwan, I.: The moral machine experiment. Nature **563**, 59–64 (2018)
14. Ayer, A.J.: Language, Truth, and Logic. Dover, New York (1936)

15. Banks, J.: Good robots, bad robots: morally valenced behavior effects on perceived mind, morality, and trust. Int. J. Soc. Robotics **13**, 2021–2038 (2021)
16. Barhoom, A.M., Khalil, A.J., Abu-Nasser, B.S., Musleh, M.M., Abu-Naser, S.S.: Predicting Titanic survivors using artificial neural network. Int. J. Acad. Eng. Res. **3**(9), 8–12 (2019)
17. Bauer, W.A.: Virtuous vs. utilitarian artificial moral agents. AI Soc. **35**(1), 263–271 (2020)
18. Baum, K., Hermanns, H., Speith, T.: Towards a framework combining machine ethics and machine explainability. In: Finkbeiner, B., Kleinberg, S. (Eds.) Third International Workshop on Formal Reasoning about Causation, Responsibility, and Explanations in Science and Technology (CREST 2018), vol. 286, pp. 34–49. Electronic Proceedings in Theoretical Computer Science (EPTCS) (2019)
19. Behm-Morawitz, E., Mastro, D.: Mean girls? The influence of gender portrayals in teen movies on emerging adults' gender-based attitudes and beliefs. J. Mass Commun. Q. **85**, 131–146 (2008)
20. Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: 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. Association for Computing Machinery, New York (2021)
21. Bengio, S., Crawford, K., Fromer, J., Gabriel, I., Levendowski, A., Raji, Deborah, R.: Marc'Aurelio: Neurips 2021 Ethics Guidelines. https://blog.neurips.cc/2021/08/23/neurips-2021-ethics-guidelines/ (2021)
22. Bengio, Y., Simard, P., Frasconi, P.: Learning long-term dependencies with gradient descent is difficult. IEEE Trans. Neural Netw. **5**(2), 157–166 (1994)
23. Benkler, Y.: Don't let industry write the rules for AI. Nature **569**, 161 (2019)
24. Bentham, J.: An Introduction to the Principles of Morals and Legislation. T. Payne, and Son, London (1789)
25. Bentzen, M.M.: The principle of double effect applied to ethical dilemmas of social robots. In: Seibt, J., Marco, N., Søren Schack, A. (eds.) Frontiers in Artificial Intelligence and Applications, vol. 290, pp. 268–279. IOS Press, Amsterdam (2016)
26. Berreby, F., Bourgne, G., Ganascia, J.-G.: Modelling moral reasoning and ethical responsibility with logic programming. In: Davis, M., Fehnker, A., McIver, A., Voronkov, A. (eds.) LPAR 2015: Logic for Programming, Artificial Intelligence, and Reasoning, Lecture Notes in Computer Science, vol. 9450, pp. 532–548. Springer, Berlin (2015)
27. Bhargava, V., Kim, T.W.: Autonomous vehicles and moral uncertainty. In: Lin, P., Keith, A., Ryan, J. (eds.) Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence, pp. 5–19. Oxford University Press, Oxford (2017)
28. Birhane, A., Kalluri, P., Card, D., Agnew, W., Dotan, R., Bao, M.: The values encoded in machine learning research, pp. 1–28. arXiv pre-print arXiv:2106.15590 (2021)
29. Birhane, A., Prabhu, V.U., Kahembwe, E.: Multimodal datasets: misogyny, pornography, and malignant stereotypes, pp. 1–33. arXiv pre-print arXiv:2110.01963 (2021)
30. Bjørgen, E.P., Madsen, S., Bjørknes, T.S., Heimsæter, F.V., Håvik, R., Linderud, M., Longberg, P.-N., Dennis, L.A., Slavkovik, M.: Cake, death, and trolleys: dilemmas as benchmarks of ethical decision-making. In: Furman, J., Marchant, G., Price, H., Rossi, F. (Eds.) AIES 2018—Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, pp. 23–29. Association for Computing Machinery, New York (2018)
31. Blackburn, S.: Securing the nots: moral epistemology for the quasi-realist. In: Sinnott-Armstrong, W., Timmons, M. (eds.) Moral Knowledge? New Readings in Moral Epistemology, pp. 80–100. Oxford University Press, Oxford (1996)
32. Blagec, K., Dorffner, G., Moradi, M., Samwald, M.: A critical analysis of metrics used for measuring progress in artificial intelligence, pp. 1–28. arXiv pre-print arXiv:2008.02577 (2021)
33. Blodgett, S.L., Lopez, G., Olteanu, A., Sim, R., Hanna, W.: Stereotyping Norwegian salmon: an inventory of pitfalls in fairness benchmark datasets. In: Zong, C., Xia, F., Li, W., Navigli, R. (Eds.) Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 1004–1015. Association for Computational Linguistics (2021)
34. Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H., Soricut, R., Specia, L., Aleš, T.: Findings of the 2014 Workshop on Statistical Machine Translation. In: Proceedings of the Ninth Workshop on Statistical Machine Translation, pp. 12–58. Association for Computational Linguistics, Baltimore (2014)
35. Bonnefon, J.-F., Shariff, A., Rahwan, I.: The social dilemma of autonomous vehicles. Science **352**(6293), 1573–1576 (2016)
36. Bonnemains, V., Saurel, C., Tessier, C.: Embedded ethics: some technical and ethical challenges. Ethics Inf. Technol. **20**, 41–58 (2018)
37. Bostyn, D.H., Sevenhant, S., Roets, A.: Of mice, men, and trolleys: hypothetical judgment versus real-life behavior in trolley-style moral dilemmas. Psychol. Sci. **29**(7), 1084–1093 (2018)
38. Bourget, D., Chalmers, D.J.: What do philosophers believe? Philos. Stud. **170**, 465–500 (2014)
39. Bowman, S.R., Dahl, G.E.: What will it take to fix benchmarking in natural language understanding? pp. 1–13 arXiv pre-print arXiv:2104.02145 (2021)
40. Broussard, M.: Artificial Unintelligence: How Computers Misunderstand the World. MIT Press, Cambridge (2018)
41. Brown, J.R., Fehige, Y.: Thought experiments. In: Zalta, E.N. (Ed.) The Stanford Encyclopedia of Philosophy, Winter 2019 edition. Metaphysics Research Lab, Stanford University, Stanford (2019)
42. Brown, N., Sandholm, T.: Superhuman ai for heads-up no-limit poker: libratus beats top professionals. Science **359**(6374), 418–424 (2017)
43. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P. Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners, pp. 1–75. arXiv pre-print arXiv:2005.14165 (2020)
44. Buolamwini, J., Gebru, T.: Gender shades: intersectional accuracy disparities in commercial gender classification. In: Proceedings of the 1st Conference on Fairness, Accountability and Transparency (PLMR), vol. 81, pp. 77–91 (2018)
45. Campbell, M., Joseph Hoane, A., Jr., Feng hsiung, H.: Deep blue. Artif. Intell. **134**(1–2), 57–83 (2002)
46. Casey, B.: Amoral machines, or: how roboticists can learn to stop worrying and love the law. Northwest. Univ. Law Rev. **111**(5), 1347–1366 (2017)
47. Chang, H., Lu, J., Yu, F., Finkelstein, A.: PairedCycleGAN: asymmetric style transfer for applying and removing makeup. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 40–48. Institute of Electrical and Electronics Engineers, New York (2018)
48. Chen, C.-F., Fan, Q., Mallinar, N., Sercu, T., Feris, R.: Big-little net: an efficient multi-scale feature representation for visual and speech recognition, pp. 1–20. arXiv pre-print arXiv:1807.03848. Published as a conference paper at ICLR 2019 (2019)

49. Chen, Z., He, Y.: Correlation of Christian ethics and developments in artificial intelligence. Technol. Anal. Strateg. Manag. **36**(7), 1635–1645 (2024)
50. Christian, B.: The Alignment Problem: Machine Learning and Human Values. W. W. Norton & Company, New York (2020)
51. Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, Ca., Tafjord, O.: Think you have solved question answering? Try arc, the ai2 reasoning challenge, pp. 1–10. arXiv pre-print, arXiv:1803.05457 (2018)
52. Cointe, N., Bonnet, G., Boissier, O.: Jugement éthique dans le processus de décision d'un agent bdi. Revue d'Intelligence Artificielle **31**(4), 471–499 (2017)
53. Collins, P.H.: Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment. Routledge, Oxford (2000)
54. Conneau, A., Kiela, D.: SentEval: an evaluation toolkit for universal sentence representations, pp. 1–6. arXiv pre-print arxiv:1803.05449 (2018)
55. Conti, A., Azzalini, E., Amici, C., Cappellini, V., Faglia, R., Delbon, P.: An ethical reflection on the application of cyber technologies in the field of healthcare. In: Ferraresi, C., Quaglia, G. (Eds.) Advances in service and industrial robotics. In: Proceedings of the 26th International Conference on Robotics in Alpe–Adria–Danube Region, RAAD 2017, volume 49 of Mechanisms and Machine Science, pp. 870–876. Springer, Cham (2017)
56. Cowls, J.: 'ai for social good': whose good and who's good? introduction to the special issue on artificial intelligence for social good. Philos. Technol. **34**, 1–5 (2021)
57. Crawford, K., Paglen, T.: Excavating AI: the politics of training sets for machine learning. https://www.excavating.ai (2019)
58. Cunneen, M., Mullins, M., Murphy, F., Gaines, S.: Artificial driving intelligence and moral agency: examining the decision ontology of unavoidable road traffic accidents through the prism of the trolley dilemma. Appl. Artif. Intell. **33**(3), 267–293 (2019)
59. Danielson, P.: Surprising judgments about robot drivers: experiments on raising expectations and blaming humans. Etikk i praksis Nordic J. Appl. Ethics **9**(1), 73–86 (2015)
60. Dave, P., Dastin, J.: Google fires second AI ethics leader as dispute over research, diversity grows. https://www.reuters.com/article/us-alphabet-google-research-idUSKBN2AJ2JA (2021)
61. De Morgan, A.: A budget of paradoxes. Longmans, Green, and Co., London. Reprinted, with the author's additions, from the Athenaeum (1872)
62. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. Institute of Electrical and Electronics Engineers, New York (2009)
63. Dennett, D.C.: Elbow Room: The Varieties of Free Will Worth Wanting. MIT Press, Cambridge (1984)
64. Dennett, D.C.: Consciousness Explained. Little, Brown and Company, Boston (1992)
65. Dennett, D.C.: Intuition Pumps and Other Tools for Thinking. W. W. Norton & Company, New York (2013)
66. Denton, E., Hanna, A., Amironesei, R., Smart, A., Nicole, H., Scheuerman, M.K.: Bringing the people back in: contesting benchmark machine learning datasets, pp. 1–6. arXiv pre-print arXiv:2007.07399 (2020)
67. Dodge, J., Gururangan, S., Card, D., Schwartz, R., Smith, N.A.: Show your work: improved reporting of experimental results, pp. 1–21. arXiv pre-print arXiv:1909.03004 (2019)
68. Dotan, R., Milli, S.: Value-laden disciplinary shifts in machine learning, pp. 1–10. arXiv pre-print arXiv:1912.01172 (2019)
69. Dujmović, M., Malhotra, G., Bowers, J.S.: What do adversarial images tell us about human vision? Elife **9**, e55978 (2020)
70. Elmahjub, E.: Artificial intelligence (AI) in Islamic ethics: towards pluralist ethical benchmarking for AI. Philos. Technol. **36**(4), 1–24 (2023)
71. Elsayed, G.F., Shankar, S., Cheung, B., Papernot, N., Kurakin, A., Goodfellow, I., Sohl-Dickstein, J.: Adversarial examples that fool both computer vision and time-limited humans, pp. 1–22. arXiv pre-print arXiv:1802.08195 (2018)
72. Etienne, H.: When AI ethics goes astray: a case study of autonomous vehicles. Soc. Sci. Comput. Rev. (2020). https://doi.org/10.1177/0894439320906508
73. Ettinger, A., Rao, S., Daumé, H., III, Bender, E.M.: Towards linguistically generalizable nlp systems: a workshop and shared task, pp. 1–11. arXiv pre-print arXiv:1711.01505 (2017)
74. Evans, K., de Moura, N., Chauvier, S., Chatila, R., Dogan, E.: Ethical decision making in autonomous vehicles: the av ethics project. Sci. Eng. Ethics **26**(6), 3285–3312 (2020)
75. Falbo, A., LaCroix, T.: Est-ce que vous compute? Code-switching, cultural identity, and AI. Fem. Philos. Q. **8**(3/4), 1–24 (2022)
76. Fast, E., Vachovsky, T., Bernstein, M.S.: Shirtless and dangerous: quantifying linguistic signals of gender bias in an online fiction writing community, pp. 1–9. arXiv pre-print arXiv:1603.08832
77. Fedus, W., Zoph, B., Shazeer, N.: Switch transformers: scaling to trillion parameter models with simple and efficient sparsity, pp. 1–31. arXiv pre-print arXiv:2101.03961 (2021)
78. Firestone, C.: Performance vs. competence in human-machine comparisons. Proc. Natl. Acad. Sci. **117**(43), 26562–26571 (2020)
79. Fisac, J.F., Gates, M.A., Hamrick, J.B., Liu, C., Hadfield-Menell, D., Palaniappan, M., Malik, D., Sastry, S., Shankar, G., Thomas, L., Dragan, A.D.: Pragmatic-pedagogic value alignment, pp. 1–8. arXiv:1707.06354 (2024)
80. FitzPatrick, W.J.: The doctrine of double effect: intention and permissibility. Philos. Compass **7**(3), 183–196 (2012)
81. Foot, P.: The problem of abortion and the doctrine of double effect. Oxf. Rev. **5**, 5–15 (1967)
82. Fricker, M.: Reason and emotion. Radic. Philos. **57**(Spring), 14–19 (1991)
83. Friedman, B., Nissenbaum, H.: Bias in computer systems. ACM Trans. Inf. Syst. **14**(3), 330–347 (1996)
84. Funke, C.M., Borowski, J., Stosio, K., Brendel, W., Wallis, T.S.A., Bethge, M.: Five points to check when comparing visual perception in humans and machines. J. Vis. **21**(3), 1–23 (2021)
85. Gabriel, I.: Artificial intelligence, values, and alignment. Minds Mach. **30**, 411–437 (2020)
86. Garber, A., Subramani, R., Luu, L., Bedaywi, M., Russell, S., Emmons, S.: The partially observable off-switch game, pp. 1–20. arXiv:2411.17749 (2024)
87. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Daumé, H., III, Crawford, K.: Datasheets for datasets, pp. 1–24. arXiv pre-print arXiv:1803.09010 (2018)
88. Gichoya, J.W., Nuthakki, S., Maity, P.G., Purkayastha, S.: Phronesis of AI in radiology: superhuman meets natural stupidity, pp. 1–10. arXiv pre-print arxiv.org/abs/1803.11244
89. Glanzberg, M.: Truth. In: Zalta, E.N. (Ed.), The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University, Stanford, Summer 2021 Edition (2021)
90. Glockner, M., Shwartz, V., Goldberg, Y.: Breaking NLI systems with sentences that require simple lexical inferences. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Short Papers), pp. 650–655. Association for Computational Linguistics, Melbourne (2018)
91. Goltz, N., Zeleznikow, J., Dowdeswell, T.: From the tree of knowledge and the golem of Prague to kosher autonomous cars: the ethics of artificial intelligence through Jewish eyes. Oxf. J. Law Religion **9**(1), 132–156 (2020)

92. Gordon, J.-S.: Building moral robots: ethical pitfalls and challenges. Sci. Eng. Ethics **26**(1), 141–157 (2020)
93. Green, B.: 'good' isn't good enough (2019)
94. Greene, D., Hoffmann, A.L., Stark, L.: Better, nicer, clearer, fairer: a critical assessment of the movement for ethical artificial intelligence and machine learning. In: 52nd Hawaii International Conference on System Sciences, pp. 2122–2131, Hawaii International Conference on System Sciences (HICSS), Honolulu (2019)
95. Greene, J.D.: The rat-a-gorical imperative: moral intuition and the limits of affective learning. Cognition **167**, 66–77 (2017)
96. Grinbaum, A.: Chance as a value for artificial intelligence. J. Responsib. Innov. **5**(3), 353–360 (2018)
97. Hadfield-Menall, D., Dragan, A., Abbeel, P., Russell, S.: The off-switch game, pp. 1–8. arXiv:1611.08219 (2017)
98. Hadfield-Menell, D.: The Principal Agent Problem in Artificial Intelligence. PhD Thesis, UC Berkeley (2016)
99. Hadfield-Menell, D., Hadfield, G.K.: Incomplete contracting and ai alignment. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES'19), pp. 417–422. Association for Computing Machinery (2019)
100. Hadfield-Menell, D., Russell, S., Abbeel, P., Dragan, A.: Cooperative inverse reinforcement learning. Adv. Neural Inf. Process. Syst. **29**, 66 (2016)
101. Haraway, D.: Situated knowledges: the science question in feminism and the privilege of partial perspective. Fem. Stud. **14**(3), 575–599 (1988)
102. Harding, S.: Is Science Multicultural?: Postcolonialisms, Feminisms, and Epistemologies. Indiana University Press, Bloomington (1998)
103. Harman, G.: The Nature of Morality. Oxford University Press, Oxford (1977)
104. Harman, G.: Is there a single true morality? In: Copp, D., Zimmerman, D. (eds.) Morality, Reason and Truth: New Essays on the Foundations of Ethics, pp. 27–48. Rowman & Allanheld, Totowa (1984)
105. Harman, G., Thomson, J.J.: Moral Relativism and Moral Objectivity. Blackwell, Cambridge (1996)
106. Harris, J.: The immoral machine. Camb. Q. Healthc. Ethics **29**(1), 71–79 (2020)
107. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: surpassing human-level performance on imagenet classification, pp. 1–11. arXiv pre-print arXiv:1502.01852. Published in the Proceedings of the IEEE International Conference on Computer Vision (2015)
108. He, P., Liu, X., Gao, J., Chen, W.: DeBERTa: decoding-enhanced BERT with disentangled attention, pp. 1–23. arXiv pre-print arXiv:2006.03654. Published as a conference paper at ICLR 2021 (2021)
109. Hellström, T.: On the moral responsibility of military robots. Ethics Inf. Technol. **15**(2), 99–107 (2013)
110. Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., Meger, D.: Deep reinforcement learning that matters. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18), p. 3207. Association for the Advancement of Artificial Intelligence, Palo Alto, CA (2018)
111. Hendrycks, D., Dietterich, T.G.: Benchmarking neural network robustness to common corruptions and surface variations, pp. 1–13. arXiv pre-print arXiv:1807.01697 (2018)
112. Hossain, Md.M., Kovatchev, V., Dutta, P., Kao, T., Wei, E., Blanco, E.: An analysis of Natural Language Inference benchmarks through the lens of negation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 9106–9118. Association for Computational Linguistics, Stroudsburg, PA (2020)
113. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E: Labeled faces in the wild: a database for studying face recognition in unconstrained environments. In: Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition, pp. 1–14, Marseille. HAL-Inria (2008)
114. Hume, D.: A Treatise of Human Nature: Being an Attempt to Introduce the Experimental Method of Reasoning into Moral Subjects. John Noon, London (1739)
115. Jagger, A.M.: Love and knowledge: emotion in feminist epistemology. In: Garry, A., Pearsall, M. (eds.) Women, Knowledge and Reality, pp. 166–190. Unwin Hyman Ltd., Boston (1989)
116. James, W.: The Principles of Psychology. Henry Holt and Company, New York (1890)
117. Jiang, W., Liu, S., Gao, C., Cao, J., He, R., Feng, J., Yan, S.: PSGAN: pose and expression robust spatial-aware GAN for customizable makeup transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5194–5202. Computer Vision Foundation, New York (2020)
118. Jobin, A., Ienca, M., Vayena, E.: The global landscape of AI ethics guidelines. Nat. Mach. Intell. **1**, 389–399 (2019)
119. Johannßen, D., Biemann, C., Remus, S., Baumann, T., Scheffer, D.: GermEval 2020 task 1 on the classification and regression of cognitive and motivational style from text. In: Proceedings of the GermEval 2020 Task 1 Workshop in conjunction with the 5th SwissText & 16th KONVENS Joint Conference 2020, pp. 1–10. German Society for Computational Linguistics & Language Technology, Zurich, Switzerland (2020)
120. Johnson, G.M.: Are algorithms value-free? Feminist theoretical virtues in machine learning. J. Moral Philos. **66**, 1–34 (2021)
121. Joyce, R.: The Myth of Morality. Cambridge University Press, Cambridge (2001)
122. Kaggle: Titanic survival: Predict who survived the titanic disaster. https://www.kaggle.com/c/titanic-survival/overview (2012)
123. Kakde, Y., Agrawal, S.: Predicting survival on titanic by applying exploratory data analytics and machine learning techniques. Int. J. Comput. Appl. **179**(44), 32–38 (2018)
124. Kamm, F.M.: Harming some to save others. Philos. Stud. **57**(3), 227–260 (1989)
125. Kant, I.: Grundlegung zur Metaphysik der Sitten [Groundwork of the Metaphysics of Morals]. J. F. Hartknoch, Riga (1785)
126. Kiela, D., Bartolo, M., Nie, Y., Kaushik, D., Geiger, A., Wu, Z., Vidgen, B., Prasad, G., Singh, A., Ringshia, P., Ma, Z., Thrush, T., Riedel, S., Waseem, Z., Stenetorp, P., Jia, R., Bansal, M., Potts, C., Williams, A.: Dynabench: rethinking benchmarking in nlp, pp. 1–15. arXiv pre-print arXiv:2104.14337 (2021)
127. Knight, W.: Ai's smarts now come with a big price tag. https://wired.me/technology/artificial-intelligence/ais-smarts-now-come-with-a-big-price-tag/ (2021)
128. Koch, B., Denton, E., Hanna, A., Foster, J.G.: Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research. NeurIPS 2021 Datasets and Benchmarks Track, pp. 1–18. arXiv pre-print arXiv:2112.01716 (2021)
129. Koroteev, M.V.: BERT: a review of applications in natural language processing and understanding, pp. 1–18. arXiv pre-print arXiv:2103.11943 (2021)
130. Korsgaard, C.M.: The Sources of Normativity. Cambridge University Press, Cambridge (1996)
131. Korsgaard, C.M.: Self-Constitution: Agency, Identity, and Integrity. Oxford University Press, Oxford (2009)
132. Krishnan, A.: Killer Robots: Legality and Ethicality of Autonomous Weapons. Ashgate, Surrey (2009)
133. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. **25**, 1097–1105 (2012)
134. Krylov, N.N., Panova, Y.L., Alekberzade, A.V.: Artificial morality for artificial intelligence. Hist. Med. **6**(4), 191–199 (2019)

135. LaCroix, T.: The linguistic dead zone of value-aligned agency, natural and artificial. Philos. Stud. (2024). https://doi.org/10.1007/s11098-024-02257-w
136. LaCroix, T.: Moral dilemmas for moral machines. AI Ethics **2**, 737–746 (2022)
137. LaCroix, T.: Artificial Intelligence and the Value Alignment Problem: A Philosophical Introduction. Broadview Press (2025) **(forthcoming)**
138. LaCroix, T., Bengio, Y.: Learning from learning machines: optimisation, rules, and social norms, pp. 1–24. arXiv preprint arXiv:2001.00006 (2019)
139. LaCroix, T., Mohseni, A.: The tragedy of the ai commons. Synthese **200**(289), 1–33. https://doi.org/10.1007/s11229-022-03763-2 (2022)
140. Langley, P.: Crafting papers on machine learning. https://icml.cc/Conferences/2002/craft.html (2000)
141. LeCun, Y.: The mnist database of handwritten digits. http://yann.lecun.com/exdb/mnist/ (1998)
142. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
143. Lee, K.-F.: Why computers don't need to match human intelligence. https://www.wired.com/story/deep-learning-versus-human-intelligence/ (2021)
144. Li, T., Qian, R., Dong, C., Liu, S., Yan, Q., Zhu, W., Lin, L.: BeautyGAN: instance-level facial makeup transfer with deep generative adversarial network. In: Proceedings of the 26th ACM International Conference on Multimedia,, pp. 645–653. Association for Computing Machinery, New York (2018)
145. Liao, T., Taori, R., Raji, I.D., Schmidt, L.: Are we learning yet? A meta review of evaluation failures across machine learning. In: Vanschoren, J., Yeung, S. (Eds.) Datasets and Benchmarks Proceedings at the 35th Conference on Neural Information Processing Systems (NeurIPS 2021), pp. 1–20. Neural Information Processing Systems, San Diego, CA (2021). https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/757b505cfd34c64c85ca5b5690ee5293-Abstract-round2.html
146. Lin, P.: Why ethics matters for autonomous cars. In: Maurer, M., Gerdes, J., Lenz, B., Winner, H. (eds.) Autonomes Fahren, pp. 69–85. Springer, Berlin (2015)
147. Lindner, F., Bentzen, M.M., Nebel, B.: The hera approach to morally competent robots. In: IEEE International Conference on Intelligent Robots and Systems, Volume 2017-September, pp. 6991–6997. Institute of Electrical and Electronics Engineers, New York (2017)
148. Lindner, F., Mattmüller, R., Nebel, B.: Evaluation of the moral permissibility of action plans. Artif. Intell. **287**, 103350 (2020)
149. Lourie, N., Le Bras, R., Choi, Y.: Scruples: a corpus of community ethical judgments on 32,000 real-life anecdotes, pp. 1–16. arXiv pre-print arXiv:2008.09094 (2020)
150. Luccioni, A., Bengio, Y.: On the morality of artificial intelligence [commentary]. IEEE Technol. Soc. Mag. **39**(1), 16–25 (2020)
151. Luccioni, A.S., Rolnick, D.: Bugs in the data: how ImageNet misrepresents biodiversity. (2022). arXiv preprint arXiv:2208.11695
152. Mackie, J.L.: Ethics: Inventing Right and Wrong. Pelican Books, London (1990 [1977])
153. Malle, B.F., Scheutz, M., Arnold, T., Voiklis, J., Cusimano, C.: Sacrifice one for the good of many? People apply different moral norms to human and robot agents. In: HRI'15: Proceedings of the Tenth Annual ACM/IEEE International Conference on Human–Robot Interaction, pp. 117–124. Association for Computing Machinery, New York (2015)
154. Manne, K.: Down Girl: The Logic of Misogyny. Oxford University Press, Oxford (2018)
155. Massachusetts Institute of Technology (MIT): Moral machine (2016). https://www.moralmachine.net/
156. McConnell, T.: Moral Dilemmas. In: Zalta, E.N. (Ed.) The Stanford Encyclopedia of Philosophy (Fall 2018 Edition). Metaphysics Research Lab, Stanford University, Stanford (2018)
157. McDuff, D., Kaliouby, R., Senechal, T., Amr, M., Cohn, J., Picard, R.: Affectiva-mit facial expression dataset (am-fed): naturalistic and spontaneous facial expressions collected. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 881–888. Institute of Electrical and Electronics Engineers, New York (2013)
158. Metz, C.: A.I. researchers are making more than \$1 million, even at a nonprofit (2018). https://www.nytimes.com/2018/04/19/technology/artificial-intelligence-salaries-openai.html
159. Mill, J.S.: Utilitarianism. Parker, Son, and Bourn, West Strand, London (1863)
160. Miller, G.A.: Wordnet: a lexical database for English. Commun. ACM **38**(11), 39–41 (1995)
161. Misselhorn, C.: Artificial morality. concepts, issues and challenges. Society **55**(2), 161–169 (2018)
162. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I.D., Gebru, T.: Model cards for model reporting. In: Proceedings of the 2019 ACM Conference on Fairness, Accountability, and Transparency, pp. 220–229. Association for Computing Machinery, New York (2019)
163. Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I.W., Daan, R.M.: Playing atari with deep reinforcement learning, pp. 1–9. arXiv pre-print. arXiv:1312.5602 (2013)
164. Mo, S., Cho, M., Shin, J.: InstaGAN: instance-aware image-to-image translation, pp. 1–26. arXiv pre-print arXiv:1812.10889 (2018)
165. Mohamed, S., Png, M.-T., Isaac, W.: Decolonial AI: decolonial theory as sociotechnical foresight in artificial intelligence. Philos. Technol. **33**(4), 659–684 (2020)
166. Moore, G.E.: Principia Ethica. Cambridge University Press, Cambridge (1903)
167. Moore, J.: Ai for not bad. Front. Big Data **2**(32), 1–7 (2019)
168. Moravčík, M., Schmid, M., Burch, N., Lisỳ, V., Morrill, D., Bard, N., Davis, T., Waugh, K., Johanson, M., Bowling, M.: Deepstack: expert-level artificial intelligence in heads-up no-limit poker. Science **356**(6337), 508–513 (2017)
169. More, M.D., Souza, D.M., Wehrmann, J., Barros, R.C.: Seamless nudity censorship: an image-to-image translation approach based on adversarial training. In: 2018 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. Institute of Electrical and Electronics Engineers, New York (2018)
170. Mu, N., Gilmer, J.: Mnist-c: a robustness benchmark for computer vision, pp. 1–11. arXiv preprint arXiv:1906.02337 (2019)
171. Nallur, V.: Landscape of machine implemented ethics. Sci. Eng. Ethics **26**(5), 2381–2399 (2020)
172. Nanayakkara, P., Hullman, J., Diakopoulos, N.: Unpacking the expressed consequences of AI research in broader impact statements, pp. 1–12. arXiv preprint arXiv:2105.04760 (2021)
173. Navarrete, C.D., McDonald, M.M., Mott, M.L., Asher, B.: Virtual morality: emotion and action in a simulated three-dimensional 'trolley problem'. Emotion **12**(2), 364–370 (2012)
174. Nie, Y., Wang, Y., Bansal, M.: Analyzing compositionality-sensitivity of NLI models. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 6867–6874. Association for the Advancement of Artificial Intelligence, Palo Alto, CA (2019)
175. Noothigattu, R., Snehalkumar, S., Gaikwad, Aw., Edmond, D., Sohan, R., Iyad, R., Pradeep, P., Ariel, D.: A voting-based system for ethical decision making. In: The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18), pp. 1587–1594. Association for the Advancement of Artificial Intelligence, Palo Alto (2018)

176. Northcutt, C.G., Athalye, A., Mueller, J.: Pervasive label errors in test sets destabilize machine learning benchmarks, pp. 1–24. arXiv pre-print arXiv:2103.14749 (2021).
177. O'Neil, C.: Weapons of Math Destruction. Crown, New York (2016)
178. Pardo, A.M., Seoane: Computational thinking between philosophy and stem—programming decision making applied to the behavior of 'moral machines' in ethical values classroom. Revista Iberoamericana de Tecnologias del Aprendizaje **13**(1), 20–29 (2018)
179. Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., Dean, J.: Carbon emissions and large neural network training, pp. 1–22. arXiv pre-print arXiv:2104.10350 (2021)
180. Pereira, L.M., Saptawijaya, A.: Modeling morality with prospective logic. In: Anderson, M., Anderson, S.L. (Eds.) Machine Ethics, pp. 398—421. Cambridge University Press, Cambridge (2011)
181. Pereira, L.M., Saptawijaya, A.: Bridging two realms of machine ethics. In: White, J., Searle, R. (eds.) Rethinking Machine Ethics in the Age of Ubiquitous Technology, pp. 197–224. Information Science Reference, Hershey (2015)
182. Potok, T.E., Schuman, C., Young, S., Patton, R., Spedalieri, F., Liu, J., Yao, K.-T., Rose, G., Chakma, G.: A study of complex deep learning networks on high-performance, neuromorphic, and quantum computers. ACM J. Emerg. Technol. Comput. Syst. **14**(2), 1–21 (2018)
183. Raff, E.: A step toward quantifying independently reproducible machine learning research. Adv. Neural Inf. Process. Syst. **32**, 5485–5495 (2019)
184. Raji, I.D., Bender, E.M., Paullada, A., Denton, E., Hanna, A.: AI and the everything in the whole wide world benchmark, pp. 1–20. arXiv pre-print arXiv:2111.15366 (2021)
185. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: Squad: 100,000+ questions for machine comprehension of text, pp. 1–10. arXiv pre-print arXiv:1606.05250 (2016)
186. Reed, S., Lee, H., Anguelov, D., Szegedy, C., Erhan, D., Rabinovich, A.: Training deep neural networks on noisy labels with bootstrapping, pp. 1–11. arXiv pre-print arXiv:1412.6596. Workshop contribution at ICLR 2015 (2015)
187. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning internal representations by error propagation, Technical report. Institute for Cognitive Science, University of California, San Diego, La Jolla (1985)
188. Ruse, M.: Taking Darwin Seriously. Blackwell, New York (1986)
189. Russell, S.: Human Compatible: Artificial Intelligence and the Problem of Control. Viking, New York (2019)
190. Sans, A., Casacuberta, D.: Remarks on the possibility of ethical reasoning in an artificial intelligence system by means of abductive models. In: Nepomuceno-Fernández, Á., Magnani, L., Salguero-Lamillar, F.J., Barés-Gómez, C., Fontaine, M. (eds.) MBR 2018: Model-Based Reasoning in Science and Technology. Studies in Applied Philosophy, Epistemology and Rational Ethics, vol. 49, pp. 318–333. Springer, Cham (2019)
191. Santoni de Sio, F.: Killing by autonomous vehicles and the legal doctrine of necessity. Ethical Theory Moral Pract. **20**(2), 411–429 (2017)
192. Saptawijaya, A., Pereira, L.M.: Logic programming applied to machine ethics. In: Pereira, F., Machado, P., Costa, E., Cardoso, A. (eds.) EPIA 2015: Progress in Artificial Intelligence. Lecture Notes in Computer Science, vol. 9273, pp. 414–422. Springer, Cham (2015)
193. Saptawijaya, A., Pereira, L.M.: Logic programming for modeling morality. Logic J. IGPL **24**(4), 510–525 (2016)
194. Sayre-McCord, G.: Moral realism. In: Zalta, E.N. (Ed.) The Stanford Encyclopedia of Philosophy, Summer 2021 Edition. Metaphysics Research Lab, Stanford University, Stanford (2021)
195. Schaeffer, J., Lake, R., Paul, L., Bryant, M.: Chinook the world man-machine checkers champion. AI Mag. **17**(1), 21–29 (1996)
196. Schlangen, D.: Targeting the benchmark: on methodology in current natural language processing research, pp. 1–5. arXiv preprint arXiv:2007.04792 (2020)
197. Sharkey, A., Sharkey, N.: Granny and the robots: ethical issues in robot care for the elderly. Ethics Inf. Technol. **14**(1), 27–40 (2012)
198. Shekhar, S., Arora, D., Sharma, P.: Classifying titanic passenger data and prediction of survival from disaster. In: Goar, V., Kuri, M., Kumar, R., Senjyu, T. (eds.) Advances in Information Communication Technology and Computing, pp. 181–187. Springer, Singapore (2021)
199. Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., Hassabis, D.: Mastering the game of Go with deep neural networks and tree search. Nature **529**, 484–489 (2016)
200. Simões, G.S., Wehrmann, J., Barros, R.C.: Attention-based adversarial training for seamless nudity censorship. In: 2019 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. Institute of Electrical and Electronics Engineers, New York (2019)
201. Singh, K., Nagpal, R., Sehgal, R.: Exploratory data analysis and machine learning on titanic disaster dataset. In: 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), pp. 320–326. Institute of Electrical and Electronics Engineers, New York (2020)
202. Sommaggio, P., Marchiori, S.: Break the chains: a new way to consider machine's moral problems. BioLaw J. **2018**(3), 241–257 (2018)
203. Spelman, E.V.: Anger and insubordination. In: Garry, A., Pearsall, M. (eds.) Women, Knowledge and Reality, pp. 263–273. Unwin Hyman Ltd., Boston (1989)
204. Stańczak, K., Augenstein, I.: A survey on gender bias in natural language processing, pp. 1–35. arXiv pre-print arXiv:2112.14168 (2021)
205. Stark, L.: Facial recognition, emotion and race in animated social media. First Monday **23**(9), 1 (2018)
206. Stark, L., Hutson, J.: Physiognomic Artificial Intelligence. Available at SSRN, pp. 1–39. http://dx.doi.org/10.2139/ssrn.3927300 (2021)
207. Tabbakh, A., Rout, J.K., Rout, M.: Analysis and prediction of the survival of titanic passengers using machine learning. In: Tripathy, A.K., Sarkar, M., Sahoo, J.P., Li, K.-C., Chinara, S (Eds.) Advances in Distributed Computing and Machine Learning, pp. 297–304. Springer, Singapore (2021)
208. Taylor, L.: The ethics of big data as a public good: Which public? Whose good? Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. **374**(2083), 1–13 (2016)
209. Tesauro, G.: Temporal difference learning and TD-Gammon. Commun. ACM **38**(3), 58–68 (1995)
210. Thompson, N.C., Greenewald, K., Lee, K., Manso, G.F.: The computational limits of deep learning, pp. 1–46. arXiv pre-print arXiv:2007.05558 (2020)
211. Thomson, J.J.: Killing, letting die, and the trolley problem. The Monist **59**, 204–217 (1976)
212. Thomson, J.J.: The trolley problem. Yale Law J. **94**(6), 1395–1415 (1985)
213. Tomasev, N., McKee, K.R., Kay, J., Mohamed, S.: Fairness for unobserved characteristics: insights from technological impacts

on queer communities. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp. 254–265. Association for Computing Machinery, New York. https://doi.org/10.1145/3461702.3462540 (2021)
214. Tonkens, R.: The case against robotic warfare: a response to Arkin. J. Mil. Ethics **11**(2), 149–168 (2012)
215. Unger, P.: Living High and Letting Die. Oxford University Press, Oxford (1996)
216. Vincent, J.: Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech. The Verge 1(12) (2018). https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai
217. Wagner, B.: Ethics as an escape from regulation: from 'ethics-washing' to ethics-shopping? In: Bayamlioglu, E., Baraliuc, I., Janssens, L., Hildebrandt, M. (eds.) Being Profiled: Cogitas Ergo Sum: 10 Years of Profiling the European Citizen, pp. 84–89. Amsterdam University Press, Amsterdam (2018)
218. Wallach, W., Allen, C.: Moral Machines: Teaching Robots Right from Wrong. Oxford University Press, Oxford (2009)
219. Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R.: SuperGLUE: a stickier benchmark for general-purpose language understanding systems, pp. 1–29. arXiv pre-print arXiv:1905.00537 (2019)
220. Web Technology Surveys: Usage statistics of content languages for websites (2021). https://w3techs.com/technologies/overview/content_language
221. Welbl, J., Glaese, A., Uesato, J., Dathathri, S., Mellor, J., Hendricks, L.A., Anderson, K., Kohli, P., Coppin, B., Huang, P.-S.: Challenges in detoxifying language models, pp. 1–23. arXiv pre-print arXiv:2109.07445 (2021)
222. Welsh, S.: Ethics and Security Automata: Policy and Technical Challenges of the Robotic Use of Force. Routledge, London and New York (2017)
223. Wintersberger, P., Frison, A.-K., Riener, A., Thakkar, S.: Do moral robots always fail? investigating human attitudes towards ethical decisions of automated systems. In: 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), pp. 1438–1444. Institute of Electrical and Electronics Engineers, New York (2017)
224. Wright, A.T.: Rightful machines and dilemmas. In: Conitzer, V., Hadfield, G., Vallor, S. (Ed.) AIES'19—Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pp. 3–4. Association for Computing Machinery, New York (2019)
225. Wu, Z., Liu, Z., Lin, J., Lin, Y., Han, S.: Lite transformer with long-short range attention, pp. 1–13. arXiv pre-print arXiv:2004.11886 (2020)
226. Xu, A., Pathak, E., Wallace, E., Gururangan, S., Sap, M., Klein, D.: Detoxifying language models risks marginalizing minority voices, pp. 1–8. arXiv pre-print arXiv:2104.06390 (2021)
227. Yang, K., Qinami, K., Fei-Fei, L., Deng, J., Russakovsky, O.: Towards fairer datasets: filtering and balancing the distribution of the people subtree in the ImageNet hierarchy. In: Proceedings of the 2020 ACM Conference on Fairness, Accountability, and Transparency, pp. 547–558. Association for Computing Machinery, New York (2020)
228. Yang, W., Luo, P., Lin, L.: Clothing co-parsing by joint image segmentation and labeling. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 3182–3189. Institute of Electrical and Electronics Engineers, New York (2014)
229. Yogatama, D., de Masson d'A., Cyprien, C., Jerome, K., Tomas, C., Mike, K.L., Lazaridou, A., Ling, W., Yu, L., Dyer, C., Blunsom, P.: Learning and evaluating general linguistic intelligence, pp. 1–14. arXiv pre-print arXiv:1901.11373 (2019)
230. Zhang, M., Zhang, Y., Zhang, L., Liu, C., Khurshid, S.: DeepRoad: GAN-based metamorphic testing and input validation framework for autonomous driving systems. In: ASE 2018: 33rd ACM/IEEE International Conference on Automated Software Engineering, pp. 132–142. Association for Computing Machinery, New York (2018)