A Roadmap for the Value-Loading Problem

Lê Nguyên Hoang

September 5, 2018

Abstract

We analyze the *value-loading problem*. This is the problem of encoding moral values into an AI agent interacting with a complex environment. Like many before, we argue that this is both a major concern and an extremely challenging problem. Solving it will likely require years, if not decades, of multidisciplinary work by teams of top scientists and experts. Given how uncertain the timeline of human-level AI research is, we thus argue that a pragmatic partial solution should be designed as soon as possible.

To this end, we propose a preliminary research program. This roadmap identifies several key steps. We hope that this will allow scholars, engineers and decision-makers to better grasp the upcoming difficulties, and to foresee how they can best contribute to the global effort.

"Solving the value-loading problem is a research challenge worthy of some of the next generation's best mathematical talent." Nick Bostrom (2014).

1 Introduction

As AI is being deployed, concerns have been raised about the possible side effects of its implementation, e.g. in terms of fairness, privacy, filter bubbles, job displacement or even existential risks [10, 52, 60]. In this paper, we introduce a research program to robustly encode moral values in AIs. Our purpose is to lay the groundwork to reliably guarantee that large-scale AIs will not behave in an immoral manner. While much of our work is speculative, we believe that several of the proposed ideas will be critical to guarantee AI safety and alignment. More importantly, we hope that this will be a useful roadmap for both AI experts and non-experts to better estimate how they can best contribute to the effort.

This paper is divided into three sections. In Section 2, we shall discuss the importance of the value-loading problem. In particular, we argue that *today's best way to do good may be to work on AI safety*, with a few short-term existential-risk exceptions like biological or war risks. In Section 3, we present the main technical challenges. We also propose a roadmap to solve them. Finally, in Section 4, we shall discuss non-technical challenges, which include funding, training, lobbying, communication, collaborative work, and so on.

2 Moral for AI as a priority

Debates about AI safety are unfortunately extremely polarized. Much focus has been given to extreme views. This is unhelpful. Arguably, much of the disagreement is due to the great uncertainty about the future and about the pace of AI progress. In this section, based on experts' opinions and further considerations, we shall not argue that human-level AI is imminent. Instead, we shall argue that *it is unreasonable to completely discard the possibility that AI might reach human-level within a decade.* In fact, it seems reasonable to assign at least a 1% probability on the fact that human-level AI might be there by 2025.

This is not much. However, since this paper is about safety, *we must not base our reasoning on the average or median prediction.* Safety is about worst or near-worst case. And a 1% probability is definitely hugely concerning. This is why we argue that, even if human-level AI is not very likely in the near future, solving AI safety still is a priority. This is particularly true given that any partial solution will almost surely take years to set up.

2.1 Predictions about AI progress

There are many disagreements on AI progress. There are even disagreements on what the experts believe about AI progress. Yet, data on experts' opinions have been collected both in [10] and [21]. The predictions about AI reaching human-level[1] are graphically represented by Figure 1.

Clearly, experts disagree. Some believe that AI will almost surely reach human-level within 25 years, while others believe that it is unlikely to do so by the end of the century. It is thus presomptuous to make any definite claim on AI progress, nor on what the experts claim about AI progress.

Having said this, there are important take-aways. First, while many experts believe that AI will not reach human-level within this century, the majority of them seems to believe that it will. In fact, according to most AI experts, it seems relatively likely that the youngest among us will be living with human-level AI at some point.

Second, and more importantly for our purpose here, it is notable that a majority of AI researchers assign a non-negligible probability to human-level AI within a decade. In fact, in the 2012 survey of [10], the median AI expert assigns a 10% probability to human-level AI by 2022. Similarly, the 2016 "aggregate forecast" by [21] assigns a 10% probability to human-level AI by 2025.

This is extremely worrying. Human-level AI should be able to better understand intelligence than AI researchers, which means that it could self-improve at a much faster rate than the rate at which current AI research progresses. As a result, there seems to be a nonnegligible probability that we may observe an "intelligence explosion" within a decade.

[1] The survey of [21] defines human-level AI as "when unaided machines can accomplish every task better and more cheaply than human workers".

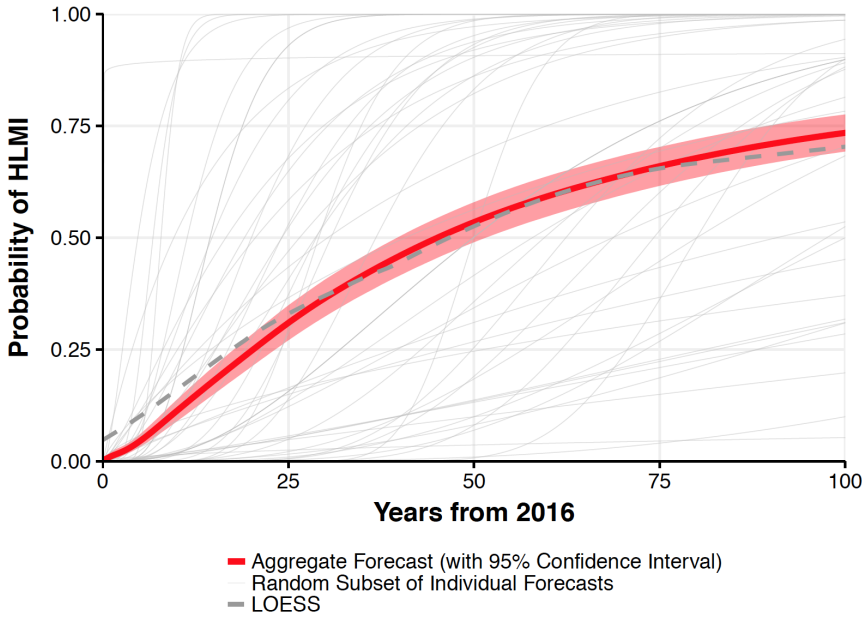

Figure 1: Experts' predictions on human-level AI from [21].

Now, evidently, experts are not entirely reliable. AI research is still shrowded with the souvenir of failed predictions of the past. Back in 1970, Marvin Minsky asserted: "In from three to eight years we will have a machine with the general intelligence of an average human being." He turned out to be deeply mistaken.

However, while there were periods during which AI experts definitely underestimated the challenges of AI research, this overoptimism should not be generalized. In fact, it seems that, lately, AI researchers have rather been underestimating the pace of AI progress. Indeed, in [21], the median AI expert predicted that AIs would need another 12 years to reach human-level at the game of Go. Again, AI experts were deeply mistaken, as AlphaGo reached human-level only a few months after the survey.

Besides, as explained by the author of the survey Katja Grace [65], AI experts do not seem to have spent much time thinking about their answers. Indeed, the mere way the question was phrased induced a bias in AI experts' answers. The survey [21] should thus not be regarded as very reliable.

But such failures of AI experts should not make us more confident in the fact that AI will not reach human-level in a near future. Au contraire, they should be regarded as an added uncertainty. But added uncertainty, e.g. larger variance, usually increases the probability of extreme scenarios. Thus, arguably, based on the survey and on experts' biases, we should assign at least a 10% probability on human-level AI by 2025. If not sooner.

2.2 Accelerating progress of AIs

Discussions about AI progress date back to the early days of the theory of computing. Stanislaw Ulam quoted John von Neumann saying that "the accelerating progress of technology [...] gives the appearance of approaching some essential singularity in the history of the race beyond which human affairs, as we know them, could not continue" [62]. This observation has been echoed by several scholars, including I.J. Good, Vernor Vinge [63], and other more recent personalities like Ray Kurzweil, Stephen Hawking, Elon Musk and Bill Gates.

The basis of their argument is simple. Technologies are critical to make better technologies. As a result, better technologies allow for faster technological progress, which allows for even better technologies, which allow for even faster technological progress, and so on. This phenomenon is well-known to be captured by Moore's law [42] in hardware development. But it should not be regarded as specific to hardware. The observation is more general. Better technologies accelerate technological progress, even in, say, software development (think about TensorFlow for instance). And this seems to inevitably lead to an exponential (if not superexponential) progress.

Unfortunately, we humans are arguably quite poor at having an intuitive understanding of such exponential growths. It seems that we tend to underestimate its effect on a longer term. This is evidenced by the fact that tales about exponential growth, such as the "wheat and chessboard problem", are often said to be extremely counter-intuitive. Our inability to foresee the speed of exponential growth should count as a warning about our inability to foresee that of technology progress.

It is often argued that a corollary of such observations is that technological progress cannot go on forever. This is precisely why some argue that it will reach some "singularity". Having said this, it seems that the main ultimate hurdle to technological progress is to be found in the laws of physics. Yet, technologies are arguably still extremely far from the limits of physics. There is still a lot of room for improvement. For one thing, we do know of a rather limited machine that performs at human-level, namely, the human brain. It seems unreasonable to claim with absolute certainty that a large-scale machine that is optimized for intelligence and computation will never outperform the human brain.

2.3 The complexity of AIs

Now, assuming that human-level AI will eventually come to being, when will this happen? To have a more informed guess of a likely answer to this question, it is useful to go back to Turing's 1950 seminal paper that started artificial intelligence research [61].

Turing argued that the main bottleneck to human-level AI was the programming of the complexity of intelligence. He postulated that the complexity of human-level intelligence was likely of the same order of magnitude as the number of synapses in the human brain. 1950 estimates of this number were between 10^{10} and 10^{15}. Turing then argued that, as a result, perhaps 10^9 bits of

Figure 2: Progress in using GAN to create *ex nihilo* realistic pictures.

program were critical to reach human-level intelligence. These days, the number of synapses is rather estimated to be around 10^{15} [13]. This should increase our estimate of the needed size of AIs to reach human-level.

As of 2017, it seems that the largest AI systems were around 10^{11} bits long [55]. This means that we are not that far from AIs whose complexity is comparable to the human brain's. Given the current pace of progress, it seems quite believable that by 2025, the complexity of AIs will exceed that of the human brain — perhaps by far. By then, AIs will also likely be able to access much more data than our human brains could ever hope to process within a lifetime. Thus, an argument similar to Turing's seems to suggest that human-level AI may be reachable by 2025.

2.4 Recent progress

In the last few years, AI research has yielded numerous outstanding results that have often been labeled as "surprising", including by AI experts. Most notable is AlphaGo's breakthrough in the game of Go. But in fact, the list of unexpected progresses in AI is itself surprisingly long. Let us only mention a few of them.

Figure 2 displays the progress in image synthesis. In particular, [32] have been able to produce images that are now hard to distinguish from actual photos. Their technology relies on so-called generative adversarial networks (GANs), which allow for the better analysis and understanding of high-dimensional inputs.

Another spectacular advance was that of Google Duplex [35], in 2018. Google Duplex is an assistant that can call and make reservations for haircut or restaurants. Its performances are hugely impressive. They are arguably indistinguishable from a competent human assistant.

Finally, let us mention the fact that the creation and optimization of seed AIs has been automated by [68] in 2017. In some sense, we thus already have self-improving AIs, as these AIs are able to solve problems by creating AIs that are improvements of themselves. In particular, the automated AI design of [68] outperforms humans' designs of AIs.

Such advances in AI are noteworthy because, according to many, they are

surprising. This means that, if AI experts had to guess how long it would take for these AIs to outperform humans at their respective tasks before finding out that they already could, most of them would have been wrong. Such AI experts would have been underestimating AI progress.

Bayesian inference tells us that, as a result, we should update our beliefs. Everytime we are surprised by AI progress, we should decrease our prediction of the date of human-level AI emergence. The fact that the last few years yielded a large number of surprising AI breakthroughs means that we should expect human-level AI sooner than we did a few years ago.

2.5 Concrete problems in AI safety

This paper is evidently not the first to raise concern about AI safety. However, much of the concern so far seems to mostly boil down to fault tolerance. In particular, [1] list 5 imaginable failures that AIs should be taught to be robust against. These are *negative side effects*, *reward hacking*, *scalable oversight*, *safe exploration* and *distributional shift*. Similar notions are also discussed in [34], who further propose environments to test pratical solutions.

The failures studied by [1] mostly concern aspects of an AI agent that the designer may not have anticipated. This is a general principle. It is often extremely hard to foresee how a system will be behaving in practice. This quote by Turing [61] is particularly relevant to understand this:

> The view that machines cannot give rise to surprises is due, I believe, to a fallacy to which philosophers and mathematicians are particularly subject. This is the assumption that as soon as a fact is presented to a mind all consequences of that fact spring into the mind simultaneously with it. It is a very useful assumption under many circumstances, but one too easily forgets that it is false. A natural consequence of doing so is that one then assumes that there is no virtue in the mere working out of consequences from data and general principles.

We should not overestimate our ability to foresee the aftermaths of launching a human-level AI.

2.6 Control is insufficient

On the other hand, [49] have proposed to allow for safe interruption of misbehaving AIs. This is a difficult problem, since an AI may learn when it is likely to be interrupted to avoid (or aim at) interruptions. Note that its desire to be, or to avoid being, interrupted could be motivated by its optimizing goal. Typically, if interruptions drive it away from rewards, then it will want to avoid interruptions. The concept of safe interruptibility has then been generalized by [15] to interacting agents.

Another approach to interruptibility consists of reliance on the AI's uncertainty about its goal, as opposed to the interrupting agent who may better know

this goal. This setting corresponds to a so-called *off-switch games* between the interrupting agent and the AI, see e.g. [22]. Because of its uncertainty about its goal, the AI may go wrong. The challenge is for the AI to analyze correctly its uncertainty, as well as the interrupting agent's ability to take over the AI's job, so as to determine when the AI should be switched off, rather than acting according to its possibly flawed belief on the agent's utility function.

Such approaches are particularly suited for addressing the *control problem*. This is the problem of maintaining human control over AIs.

However, the solutions proposed by [49] and [15] are arguably limited. Namely, they propose to constrain the AIs' learning algorithms. Yet, especially if there is a race between competing companies or countries to construct useful AIs, such safety concerns may be regarded as too costly to be implemented. They may thus be discarded by AI designers. In other words, it seems necessary to supply other potential solutions to AI safety.

Besides, for [49, 15] to be relevant, it is necessary to assume that some (human?) agent is able to control the AI and to prevent potential harms caused by the AI. Meanwhile, [22] assume that the (human?) agent will be able to take over the AI's job if needed. In practice though, large-scale AIs such as recommender systems are processing so much data that the surveillance or replacement of such AIs is arguably unfeasible, at least by humans. What is more, if an AI becomes superintelligent, then it will likely believe it can perform a better job than the interrupting agent, especially if these interrupting agents are humans. In fact, more generally, relying on humans to guarantee AI safety is probably not a good idea.

2.7 Humans are liabilities

In many systems, humans can be argued to be the bottleneck of safety. Humans can be easily mistaken, inattentive, drunk, sleepy, angered, influenced, blackmailed and threatened, e.g. through autonomous weapons (see [46]). As a result, AI safety will likely need to avoid reliance on humans.

Note that this is not specific to AI safety. Typically, this is already assumed to hold for cryptographic security. The less we rely on humans, the safer our systems. This is why it seems crucial to design systems that can guarantee their own safety. Perhaps even *despite* human intervention.

2.8 Byzantine environments

What makes large-scale AI safety particularly challenging is the fact that the environment of such AIs is extremely complex, dynamic and often adversarial. In particular, if an AI interacts with a malicious environment, this environment might make the AI choose morally bad decisions. See e.g. [51].

Much research has already been done in this direction, under the name of *adversarial learning* [38, 59]. One problem faced by this line of work is robustness to so-called *evasion attacks*. While neural networks often classify

correctly most inputs, there is often an imperceptible modification to a correctly-classified input that will turn it into a misclassified one [7, 18]. This may have major consequences, e.g. if a terrorist uses such attacks to escape surveillance.

Another problem addressed by adversarial learning is *poisoning attacks*. This is when an AI is fed with incorrectly labeled inputs during training, which may be provided by some adversary. If the AI is interacting with its environment, and if the environment contains adversaries (which it usually does), the AI's learning may then be completely upset by a few modifications of the environment. This is discussed by [8, 40, 11].

More generally, complex environments are extremely hard to handle. As an example, YouTube's recommender system is interacting with billions of users whose behaviors are hard to predict. In such systems, it is extremely hard to know if a video suggestion is morally good or bad. This may depend on the user's background knowledge and emotional state. As an extreme example, while videos on suicide methods may be generally harmless, suggesting them to some people in some contexts is arguably morally inacceptable.

2.9 Incentives

Today's individual and moral incentives in developing and deploying AIs are huge. Massive cost reduction and greater service can be provided by companies using AIs. Thousands, if not millions, of lives can be saved by the deployment of AIs in healthcare. And the same holds for self-driving cars.

Given this, it is likely to be extremely hard to slow down the progress and deployment of AIs. What is more, any discussion that demands significant increased costs and delays will likely be mostly discarded. This will be even more so if such discussions sound unilateral. No one from country X will want to stop AI deployment while country Y is not.

This is why it is important to provide pragmatic, scalable, easy-to-implement and not-too-costly solutions for AI safety. Evidently, this poses huge additional constraints for AI safety researchers. But this is why the value-loading problem in particular really needs the "best mathematical talents".

2.10 Proxies won't do it

The reason why the value-loading problem will probably not be solved easily is because measuring the moral values of decisions is a hard problem itself, especially if they are then used as optimization variables. In particular, indicators are usually extremely bad variables to optimize [47].

This principle is brilliantly captured by what has become known as Goodhart's law: "When a measure becomes a target, it ceases to be a good measure." This problem is absolutely not specific to AIs. Think about grades in educational systems. What they are based on incentivizes students to study "for the exams", which then incentivizes teachers to teach "for the exams". Especially if teachers are then scored by their students.

Similarly, metrics like GDP, profit, Shanghai ranking, h-index or number of followers have completely upset the behavior of governments, companies, universities and Twitter users. While such metrics perhaps used to be relevant indicators, their use as targets have greatly nullified their relevance. They have led to all sorts of hacking strategies, e.g. *p-hacking* in the academic world.

The same is likely to hold for AIs. Any proxy is likely to be *hacked* by AIs in some undesired and unforeseen manner. This is why it is of the utmost importance that AIs be given moral values that really are sufficiently consistent with what we humans roughly deem as accurate.

2.11 AIs can go terribly wrong

Many scholars have raised major concerns about existential risks posed by AIs [53, 67, 10]. The classical scenario is that of an AI with a misdirected goal, e.g. maximizing paperclips. Such an AI would then have the incentives to redirect all of mankind's economy towards the sole purpose of making paperclips. If humans become a hurdle to the AI's goal, then the AI will want to get rid of humans to accomplish its goal.

More generally, even when the AI does not purposely want to do harm, it is likely that its path towards achieving its goals will not be aligned with mankind's safety. In particular, it can be argued that for most objective functions, it will be *instrumentally* useful for the AI to have some partial objectives, such as, goal-preservation, self-protection, self-replication, self-improvement and resource acquisition.

This phenomenon is called *instrumental convergence*. Note that it is not specific to AI. If you want to do a lot of good, you too should probably take care of your health, read a lot to improve, and acquire money and influence.

This observation led [10] to dub doom the default scenario. If no effort is spent on solving or implementing value-loading, mankind may very well be doomed. Unfortunately, a large amount of efforts might also be insufficient.

2.12 Discussing moral is hard

What makes the challenge of implementing moral values especially difficult is that discussing moral is already hard enough among humans. We humans tend to quickly disagree even when we actually agree, especially when it concerns politics or religion [24].

It should be said that the value-loading problem is not about programming some *perfect* moral to AIs. Rather, it should be regarded as the problem of encoding moral values that are *good enough*. This is already an extremely difficult and challenging task. In fact, it may be even more complicated, once an AI will be approaching human-level and a greater proportion of the population feels that this AI's moral values will determine the future of mankind.

It is somehow both a curse and a blessing that not much focus is currently given to AIs' moral values. Of course, the trouble is that we lack the manpower to do this reliably, and the influential power to encourage AI developers to

encode moral values in their AIs. But on the bright side, it also means that we do not yet have to face irrational political controversies, nor administrative burdens to work and make great advance on this problem.

This is an additional reason why we should actively work on the value-loading problem as soon as possible. It is much better to come up with a ready-to-work solution while we can serenely and calmly work on it, and before this turns into a political debacle.

2.13 Value-loading would solve many other problems

Finally, it is noteworthy that solving the value-loading problem will likely solve many other problems. First, this is clearly the case in the presence of human-level AI. Indeed, a human-level AI will be more effective at solving any problem than we humans could. Thus, any problem that we care about would then be better solved by a human-level AI with our moral values, than if we tried to solve it without AI. This evidently includes poverty, world hunger, disease, existential risks, environmental problems, biodiversity, animal suffering, injustice, discrimination, privacy, and so on. In fact, as argued by [60], without human-level AI, in the long run, because of existential risks like asteroid collision or pandemics, mankind is doomed.

Now, arguably, solving the value-loading problem could still be extremely useful even in the absence of human-level AI. Indeed, this would allow the design of a system that stresses the most pressing moral issues. It would allow us to better allocate funds and resources to optimize our philanthropic actions. In some sense, solving the value-loading problem is essentially achieving the purpose of the *effective altruism* movement [57].

This is why we argue that, even if AI safety is of no concern to you, the value-loading problem nevertheless should be. In fact, with or without AI, if you want to do good, it seems that addressing the value-loading problem is perhaps the most impactful way to go.

3 A roadmap for value-loading

In this section, we shall present a roadmap to solve the value-loading problem. Unfortunately, our roadmap will be full of gaps and false good ideas. Our purpose is not to propose a definite perfect solution. We aim at presenting a sufficiently good starting point for others to build upon[2].

Our approach consists of identifying key steps in the design of an AI with moral values. For the clarity of exposition, these steps will be personified by 5 characters, called Alice, Bob, Charlie, Dave and Steve. Roughly speaking, Steve will be collecting data from the world, Dave will use these data to infer the likely states of the world, Charlie will compute the moral values of the likely

[2]Note that the Future Of Life Institute proposed another roadmap, which may contain other useful perspectives on the value-loading problem:
`https://futureoflife.org/landscape/`

states of the world, Bob will derive incentive-compatible rewards to motivate Alice to take the right decision, and Alice will optimize decision-making. This is summed up by Figure 3.

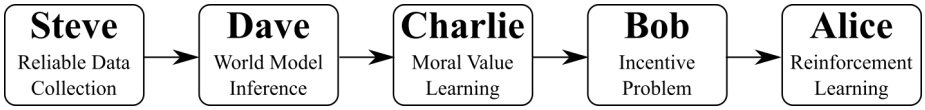

Figure 3: We propose to decompose of the value-loading problem into 5 key steps: data collection, world model inference, value learning, incentive design and reinforcement learning. We hope that such subproblems are sufficiently independent to be tackled separately.

Evidently, Alice, Bob, Charlie, Dave and Steve need not be 5 different AIs. Typically, it may be much more computationally efficient to merge Charlie and Dave. Once again, our purpose here is not to propose a ready-to-work architecture design. We aim at underlining critical steps in the design of safe and efficient AI.

3.1 Alice's Reinforcement learning

It seems that today's most promising framework for large-scale AI is that of reinforcement learning. In reinforcement learning, an AI can be regarded as a decision-making process. At time t, the AI observes some state of the world s_t. Depending on its inner parameters θ_t, it then takes (possibly randomly) some action a_t.

The decision a_t then influences the next state of the world and turns it into s_{t+1}. The transition from s_t to s_{t+1} given action a_t is usually considered to be nondeterministic. In any case, the AI then receives some reward R_t that depends on s_t, a_t and s_{t+1}. The internal parameters θ_t of the AI may then be updated into θ_{t+1} depending on previous parameters θ_t, action a_t, state s_{t+1} and reward R_t.

Note that this is a very general framework. In fact, we humans are arguably subject to this framework. At any point in time, we observe new data s_t that informs us about the world. Using an inner model of the world θ_t, we then infer what the world probably is like, which motivates us to take some action a_t. This may affect what likely next data s_{t+1} will be observed, and may be accompanied with a rewarding (or painful) feeling R_t, which will motivate us to update our inner model of the world θ_t into θ_{t+1}.

What might be more controversial is to argue that this is *all* we do. But note that this question is irrelevant to us, since we focus on moral for AIs. Not for humans. In any case, this framework seems to be the most promising framework for AIs interacting with a complex environment.

For expository purpose, let us call Alice the AI in charge of performing this reinforcement learning reasoning. Alice can thus be viewed as a maximization

algorithm, which inputs observed states s_t and rewards R_t, and undertakes actions a_t.

Such actions will probably be mostly of the form of messages sent through the Internet. This may sound benign. But it is not. If the AI is in control of 3D-printers, then a message that tells it to construct killer drones to cause a genocide would be catastrophic.

Note though that, as opposed to all other components, in some sense, Alice is the real danger. She is the only one that takes actions, in the sense that only her actions will be unconstrained (although others highly influence her decision-making and are thus critical as well).

As a result, it is of the utmost importance that Alice be well-designed. Some of the past work [49, 15] have focused on the learning algorithm, i.e. the update rule of θ_{t+1} as a function of $(\theta_t, a_t, s_{t+1}, R_t)$. However, as argued previously, it might be too costly to restrict the learning rules of AIs. Perhaps more interesting are the ideas proposed by [1] to make reinforcement learning safer, especially using *model lookahead*. This essentially corresponds to Alice simulating many likely scenarii before undertaking any action. More generally, Alice faces a *safe exploration problem*.

Even then, though, this may be insufficient, as examplified by the paperclip-maximizer example. To make sure that Alice will want to behave in a morally acceptable manner, it seems critical to at least partially control the observed state s_{t+1} and the reward R_t. Note that this is similar to the way children are taught to behave. We do so by exposing them to specific observed states, by punishing them when the sequence (s_t, a_t, s_{t+1}) is morally bad, and by rewarding them when the sequence (s_t, a_t, s_{t+1}) is morally good.

It may or may not be relevant to constrain Alice's observed state s_t. It is unclear though how to best control her by controlling what she observes, while enabling her to take reliable decisions. More work is probably needed about this. In the sequel though, we shall assume that Alice's observed state is unconstrained. In particular, we shall assume that she has access to all of Steve's data, and does inference similar to Dave's.

Whether or not Alice's observed state is constrained, her received rewards R_t are clearly critical. These rewards are her incentives, and will thus determine her desires and her decision-making. Unfortunately, determining the adequate rewards R_t to be given to Alice is an extremely difficult problem. It is, in fact, the heart of the value-loading problem. Our roadmap to solve it identifies 4 key steps incarnated by Steve, Dave, Charlie and Bob.

3.2 Steve's data collection problem

In order to do good, it is evidently crucial to be given a lot of reliable data. Indeed, even the most brilliant mind will be unable to know anything about the world if it does not have any data from that world. This is particularly true when the goal is to do good, or to make sure that one's action will not have potentially catastrophic consequences.

Evidently, much data is already available on the Internet. It is likely that any large-scale AI will have access to the Internet, as is already the case of recommender systems such as those of YouTube, Facebook, Google, Amazon or Microsoft. However, it is important to take into account the fact that the data on the Internet is not always fully reliable. It may be full of fake news, fraudulent entries, misleading videos, hacked posts and corrupted files.

It may then be relevant to invest in more reliable and relevant data collection. This would be Steve's job. Typically, Steve may want to collect economic metrics to better assess needs. Recently, it has been shown that satellite images combined with deep learning allow to compute all sorts of useful economic indicators [29], including poverty risks and agricultural productivity. It is possible that the use of still more sensors can further increase our capability to improve life standards, especially in developing countries.

To guarantee the reliability of such data, cryptographic and distributed computing solutions are likely to be useful as well, as they already are on the web. In particular, distributed computing, combined with recent Byzantine-fault-tolerant consensus algorithms like Blockchain [43] or Hashgraph [5], could guarantee the reliable storage and traceability of critical information.

Note though that such data collection mechanisms could pose major privacy issues. It is a major current challenge to balance the usefulness of collected data and the privacy violation they inevitably cause. Some possible solutions include differential privacy [14], or weaker versions like generative-adversarial privacy [28]. It could also be possible to combine these with more cryptographic solutions, like *homomorphic encryption* or *multi-party computation*. It is interesting that such cryptographic solutions may be (essentially) provably robust to any attacker, including superintelligence[3].

3.3 Dave's world model problem

Unfortunately, raw data are likely to be extremely messy, redundant, incomplete, unreliable, poisoning and even hacked. To tackle these issues, it is necessary to infer the likely actual states of the world, given the data collected by Steve. This will be Dave's job.

The overarching principle of Dave's job is probably going to be some *deep representation learning*. This corresponds to determining low-dimensional representations of high-dimensional data. This basic idea has given rise to today's most promising unsupervised machine learning alogrithms, e.g. *word vectors* [41], *autoencoders* [36] and *generative adversarial networks* (GANs) [20].

Given how crucial it is for Dave to have an unbiased representation of the world, much care will be needed to make sure that Dave's inference will foresee selection biases. For instance, when asked to provide images of CEOs, Google Image may return a greater ratio of male CEOs than the actual ratio. More generally, such biases can be regarded as instances of *Simpson's paradox* [56], and boil down to the saying "correlation is not causation". It seems crucial that

[3]The possible use of quantum computers may require postquantum cryptography.

Dave does not fall into this trap.

In fact, data can be worse than unintentionally misleading. Given how influential Alice may be, there will likely be great interest from a large number of actors to bias Steve's data gathering, and to thus fool Dave. It seems extremely important that Dave anticipate the fact that the data he was given may be purposely biased, if not hacked. Like any good journalist, Dave will likely need to cross information from different sources to infer the most likely states of the world.

This inference approach is well captured by the Bayesian paradigm [27]. In particular, Bayes rule is designed to infer the likely causes of the observed data D. These causes can also be regarded as theories T (and such theories may assume that some of the data were hacked). Bayes rule tells us that the reliability of theory T given data D can be derived formally by the following computation:

$$\mathbb{P}[T|D] = \frac{\mathbb{P}[D|T]\mathbb{P}[T]}{\mathbb{P}[D]}.$$

Importantly, Bayes rule tells us that we should not fully believe any single theory. This simply corresponds to saying that data can often be interpreted in many different mutually incompatible manners. It seems important to reason with all possible interpretations rather than isolating a single interpretation that may be flawed.

When the space of possible states of the world is large, which will surely be the case of Dave, it is often computationally intractable to reason with the full posterior distribution $\mathbb{P}[T|D]$. Bayesian methods often rather propose to sample from the posterior distribution to identify a reasonable number of good interpretations of the data. These sampling methods include Monte-Carlo methods, as well as Markov-Chain Monte-Carlo (MCMC) ones.

In some sense, Dave's job can thus be regarded as writing a compact report of all likely states of the world, given all the data collected by Steve. It is an open question as of what language Dave's report will be in. It might be useful to make it understandable by humans. But it might be too costly as well. Indeed, Dave's report might be billions of pages long. It could be unreasonable or undesirable to make it humanly readable.

One last interesting aspect to discuss is the fact that Steve and Dave are likely to gain cognitive capability over time. It is surely worthwhile to anticipate the complexification of Steve's data and of Dave's world models. It seems unclear so far how to do so. But it sounds reasonable to assume that some high-level (purely descriptive) language to describe world models is needed, and that this high-level language will have to be able to be reshaped and redesigned over time. This may be dubbed the *world description problem*. It is arguably still a very open and uncharted area of research.

As discussed above, another approach could be to more directly merge Dave's job with Charlie's.

3.4 Charlie's value learning problem

Given Dave's world models, Charlie's job will then be to compute how morally desirable the world currently is. This is known as the *value learning* problem [58]. This is the problem of determining the moral values of different states of the world. These moral values can then serve as the basis for any moral agent to take morally desirable actions.

Unfortunately, determining even a very rough approximation of what, say, the median human considers morally desirable is an extremely difficult problem. Again, it should be stressed that we should not aim at deriving an *ideal* moral, as this is likely to be a hopeless endeavor. Rather, we should try our best to make sure Charlie's values are *good enough* to avoid catastrophic outcomes, e.g. world destruction, global sufferance or major discrimination.

One proposed solution to infer human values is so-called *inverse reinforcement learning* [45, 16]. Assuming that humans perform reinforcement learning to choose their actions, and given examples of actions taken by humans in different contexts, inverse reinforcement learning infers what were the humans' likely implicit rewards that motivated their decision-making. Assuming we can somehow separate humans' selfish rewards from altruistic ones, inverse reinforcement learning seems to be a promising first step towards inferring human moral from data. There are, however, many important considerations to be taken into account, which we discuss below.

First, it is important to keep in mind that, despite Dave's effort and because of Steve's limited and possibly biased data collection, Dave's world model is fundamentally uncertain. In fact, as discussed previously, Dave would probably rather present a distribution of likely world models. Charlie's job should be regarded as a scoring of all such likely world models. In particular, she should not assign a single number to the current state of the world, but, rather, a distribution of likely values of the current state of the world. This distribution should convey the uncertainty about the actual state of the world. Besides, as we shall see, this uncertainty is likely to be crucial for Bob to choose incentive-compatible rewards for Alice adequately.

Another challenging aspect of Charlie's job will be to provide a useful representation of potential human disagreements about the moral value of the current state of the world. Human moral values are diverse and may never converge. This should not be swept under the rug. Instead, we need to agree on some way to mitigate disagreement.

Note, though, that this problem is absolutely not specific to agreement on moral values. In its most general form, this is actually a problem of *social choice*, that is, a problem of aggregating the preferences of a group of disagreeing people into a preference of the group that, in some sense, fairly well represents the individuals' preferences. Unfortunately, social choice theory is plagued with impossibility results, e.g. Arrow's theorem [3] or the Gibbard-Satterthwaite theorem [17, 54]. Again, we should not be too demanding regarding the properties of our preference aggregation. Besides, this is the path taken by social choice theory, e.g. by proposing randomized solutions to preserve some desirable

properties [25].

One particular proposal, known as *majority judgment* [6], may be of particular interest to us here. The basic idea of majority judgment is to choose some deciding quantile $q \in [0, 1]$ (often taken to be $q = 1/2$). Then, for any possible state of the world, consider all individuals' values of that state. This yields a distribution of human values for the state of the world. Majority judgment then concludes that the group's value for the state of the world is the quantile q of this distribution. If $q = 1/2$, this basically corresponds to the value chosen by the median individual of the group.

Now, to avoid an oppression of a majority over some minority, it might be relevant to choose a small value of q, say $q = 0.1$. This would mean that Charlie's assigned value to a state of the world will be less than a number v, if more than 10% of the people believe that this state is of a value less than v. But evidently, this point is very much debatable. It seems unclear so far how to best choose q.

While majority judgment seems to be a promising approach, it does raise the question of how to compare two different individuals' values. It is not clear that a value $v = 5$ given by John has a meaning comparable to Jane's $v = 5$. In fact, according to a theorem by von Neumann and Morgenstern [44], within their framework, utility functions are only defined up to a positive affine transformation. More work is probably needed to determine how to scale different individuals' utility functions appropriately, despite previous attempts in special cases [26]. Again, it should be stressed that we should not aim at an ideal solution; a workable reasonable solution is much better than no solution at all.

Now, arguably, humans' current moral values are almost surely undesirable. Indeed, over the last decades, psychology has been showing again and again that human thinking is full of inconsistencies, fallacies and cognitive biases [31]. We tend to first have a instinctive moral reaction to stories or facts [9], which quickly becomes the position we will want to defend at all costs [24]. Worse, we are unfortunately largely unaware of why we believe or want what we believe or want. This means that our current moral intuitions are unlikely to be the moral we would have, if we were more informed, thought more deeply, and tried to make sure our moral values were as well-founded as possible.

To better understand this, a thought experiment may be useful. Let us imagine better versions of us. Each *current me* is thereby associated with a me^{++}. A me^{++} is what *current me* would desire, if *current me* were smarter, thought much longer about moral, and analyzed all imaginable data of the world. Arguably, me^{++} is morally superior to *current me*. This is the fundamental claim that we should build upon.

The superiority of me^{++} over *current me* can be illustrated by the fact that past moral values are often no longer regarded as moral. Our moral intuitions of slavery, homosexuality and gender discrimination have been completely upset over the last century, if not over the last few decades. It seems unlikely that all of our other moral intuitions will never change. In particular, it seems unlikely that me^{++} will fully agree with *current me*. And it seems reasonable to argue

that me^{++} would be more right than *current me*.

These remarks motivated Eliezer Yudkowsky to introduce the concept of *coherent extrapolated volition* [66]. The basic idea is that we should adopt the moral values that future versions of ourselves would eventually adopt, if they were vastly more informed, had much more time to ponder their moral values and tried their best to be better versions of themselves. In some sense, instead of making *current me*'s debate about morals (which often turn into a pointless debacle), we should let me^{++}'s debate. In fact, since me^{++}'s supposedly already know everything about other me^{++}'s, there is actually no point in getting them to debate. It suffices to aggregate their moral values through some social choice mechanism. This is the *value aggregation problem*.

It is noteworthy that we clearly have epsitemic uncertainty about me^{++}'s. Determining me^{++}'s convictions may be called the *coherent extrapolation individual volition problem*. Interestingly, this is (mostly) a prediction problem. But it is definitely too ambitious to predict them with absolute uncertainty. Bayes rule tells us that we should rather describe these convictions by a probability distributions of likely moral values. Such values could also be approximated using a large number of proxies, as is done by *boosting methods*. The use of several proxies could avoid the overfitting of any proxy. Typically, rather than relying solely on DALYs [48], we probably should machine learning methods to combine a large number of similar metrics, especially those that aim at describing other desirable economic metrics, like human development index (HDI) or gross national happiness (GNH). Evidently, much more research is needed along these lines.

Computing the desirability of a given world state is Charlie's job. In some sense, Charlie's job would thus be to remove cognitive biases from our moral intuitions, so that they still basically reflect what we really regard as moral, but in a more coherent and informed manner. This is an incredibly difficult problem, which will likely take decades to sort out reasonably well. This is why it is of the utmost importance that it be started as soon as possible. Let us try our best to describe, informally and formally, what better versions of ourselves would likely regard as moral. Let us try to predict the volition of me^{++}'s.

This attempt is likely going to be shocking to us all. Indeed, we should expect that better versions of ourselves will find morally desirable things that the current versions of ourselves find repelling. Unfortunately though, we humans tend to react poorly to disagreeing moral values. And this is likely to hold even when the opposing moral values are our better selves'. This poses a great scientific and engineering challenge. How can one be best convinced of the moral values that he or she will eventually embrace but does not yet? In other words, how can we quickly agree with better versions of ourselves? What could someone else say to get me closer to my me^{++}? This may be dubbed the *individual moral improvement problem*.

This question is particularly critical for the value-loading problem as it will likely be a key challenge to build trust in the systems we design. But evidently, this is more general question that should be of interest to anyone who desires to do good.

3.5 Bob's incentive problem

The last piece of the jigsaw is Bob's job. Bob is in charge of computing the rewards that Alice will receive, based on the work of Steve, Dave and Charlie. Evidently he could simply compute the expectation of Charlie's assigned values for the likely states of the world. But this is probably a bad idea, as it opens the door to *reward hacking*.

It is important to keep in mind that Alice's goal is to maximize her discounted expected future rewards. But given that Alice knows (or is likely to eventually guess) how her rewards are computed, instead of undertaking the actions that we would want her to, Alice could hack Steve, Dave or Charlie's computations, so that such hacked computations yield large rewards. This is sometimes called the *wireheading problem*.

Since all this computation starts with Steve's data collection, one way for Alice to increase her rewards would be to feed Steve with fake data that will make Dave infer a deeply flawed state of the world, which Charlie may regard as ideal. Worse, Alice may then find out that the best way to do so would be to invest all of Earth's resources into misleading Steve, Dave and Charlie. This could be extremely bad for mankind.

This is why it is of the utmost important that Alice's incentives be (partially) aligned with Steve, Dave and Charlie performing well and being accurate. This will be Bob's job. Bob will need to make sure that, while Alice's rewards do convey Charlie's values, they also give Alice the incentives to make sure Steve, Dave and Charlie perform as reliably as possible the job they were given. In fact, it even seems desirable that Alice be incentivized to constantly upgrade Steve, Dave and Charlie for the better. Ideally, she would even want them to be computationally more powerful than herself, especially in the long run.

Unfortunately, it does not seem straightforward to see how Bob can make sure that Alice has such incentives. Perhaps a good idea is to penalize Dave's reported uncertainty about the likely states of the world. This source of uncertainty may have three causes. First, it may be caused by the lack of sufficiently reliable data. Bob should thus make sure Alice's rewards are affected by the reliability of Steve's data. The more reliable Steve's data, the larger Alice's rewards.

Second, it may be caused by Dave's incorrect inferences. Unfortunately, Dave's correctness may be hard to estimate. One way is to introduce an adversary AI in charge of testing Dave, as is done in Generative Adversarial Networks (GANs). But it could also be interesting to compare Dave's computations to a (computationally infeasible) baseline provided by the rigorous Bayesian computation. More generally, Dave should try to be aware of the reliability of his inferences. And the more reliable Dave's inference, the larger Alice's rewards should be.

Third, the communication between Dave and Charlie may be a bottleneck. Bob should make sure that it should be done correctly and as completely as needed. The better this communication, the larger Alice's rewards should be.

Finally, like Steve and Dave, Charlie should try to compute the uncertainty

about her computations. Again, when she feels that her estimations are unreliable, Bob should take note of this and adjust Alice's rewards accordingly to motivate Alice to provide larger resources for Charlie's computations.

Now, Bob should also mitigate the desire to retrieve more reliable data and perform more trustworthy computations with the fact that such efforts will necessarily require the exploitation of more resources, probably at the expense of Charlie's values. It is this non-trivial trade-off that Bob will need to take care of.

Bob's work might be simplified by some (partial) control of Alice's action or world model. Although it seems unclear so far how, techniques like interactive proof (IP) [4, 19] or probabilistically checkable proof (PCP) [2] might be useful to force Alice to prove its correct behavior in a computationally tractable manner. Indeed, by requesting such proofs to yield large rewards, Bob may incentivize Alice's transparency. All such considerations make up Bob's *incentive problem*.

It may or may not be useful to allow Bob to switch off Alice if need be. It should be stressed though that interruptibility is a difficult challenge, as discussed by [49, 15, 39, 22, 23, 64] among others. In fact, safe interruptibility seem to require very specific circumstances, e.g. Alice being indifferent to interruption, Alice being programmed to be suicidal in case of potential harm or Alice having more uncertainty about her rewards than Bob being able to take over Alice's job. It seems unclear so far how relevant such circumstances will be to Bob's *control problem* over Alice[4]. Besides, instead of interrupting Alice, Bob might prefer to guide Alice towards preferrable actions by acting on Alice's rewards.

On another note, it may be computationally more efficient for all if, instead of merely transmitting a reward, Bob also feeds Alice with "backpropagating signals", that is, information not about the reward itself, but about its gradient with respect to key variables, e.g. Charlie's values or Steve's reliability. Having said this, we leave open the technical question of how to best design this.

3.6 Decentralization and heuristics

We have separated the value-loading problem into 5 components for the sake of exposition. However, it is probably worthwhile to actually decompose it into many more modules to take advantage of the reliability and scalability of decentralization. In other words, instead of having a single Alice, a single Bob, a single Charlie, a single Dave and a single Steve, it seems crucial to construct multiple Alices, Bobs, Charlies, Daves and Steves.

Such a decentralization is key to fault-tolerance. Indeed, a single computer doing Bob's job could crash and leave Alice without reward nor penalty. But if Alice's rewards are an aggregate of rewards given by a large number of Bobs, then even if some of the Bobs crash, Alice's rewards will still mostly remain the same.

[4]Note though that this may be very relevant assuming that there are several Alices, as will be proposed later on.

Note though that crash-tolerance is likely to be insufficient. Instead, we should design *Byzantine-fault tolerant* mechanisms to perform the aggregation of Bobs' rewards, that is, mechanisms that still perform correctly despite the presence of hacked or malicious Bobs that would want to upset Alice's behavior. Byzantine-fault tolerance might be best guaranteed by estimators with large statistical breakdowns [37], e.g. (geometric) medians and variants [8].

Evidently, in this Byzantine environment, cryptography, especially (postquantum?) cryptographical signatures and hashes, are likely to play a critical role. Typically, Bobs' rewards will likely need to be signed, so that Alice will not be able to design fake Bobs to feed her with infinite rewards. More generally, the careful design of secure communication channels between the components of the AIs seem key. This may be called the *secure messaging problem*.

Another difficulty is the addition of more powerful and precise Bobs, Charlies, Daves and Steves to the pipeline. It is not yet clear how to best integrate reliable new comers, especially given that such new comers are likely to be constructed by Alice, and may thus be malicious. In fact, they may want to first appear benevolant to gain admission. But once they are numerous enough, they could take over the pipeline and, say, feed Alice with infinite rewards. This is the *upgrade problem*. Perhaps the code of new Bobs, Charlies, Daves and Steves should be open, and an AI should be in charge of testing these new AIs to make sure that they do not have some sort of back door created by Alice. Again, this upgrade problem seems to be an unchartered area of research.

Now, in addition to reliability, decentralization may also enable different Alices, Bobs, Charlies, Daves and Steves to focus on specific tasks. This would allow to separate different problems, which could lead to more optimized solutions. To this end, it may be relevant to adapt different Alices' rewards to their specific tasks. Note though that this could also be a problem, as Alices may enter in competition with one another like in the prisoner's dilemma. We may call it the *specialization problem*. Again, there seems to be a lot of new research needed to address this problem.

3.7 When to assign the moral burden?

Another important point to address is the extent to which interacting AIs should be made to react to Bobs' rewards. Typically, if a small company creates its own AI, should this AI be subject to our value-loading framework? It should be noted that being subject to Bobs' rewards may be computationally very demanding, as it may be hard to separate the signal of interest to the AI from the noise of Bobs' rewards.

Intuitively, the more influential this AI is, the more it should be influenced by Bobs' rewards. But even if this AI is small, it may be important to demand that it be influenced by Bobs, as, otherwise, there may be some *diffusion of responsibility*, i.e. many small AIs that disregard moral concerns on the ground that they each hardly have any global impact on the world.

Another potential challenge is the fact that an AI may gain computational capability and influence over time. If no value-loading is prepared for this

AI, it might eventually become a human-level AI with no moral values, which may be catastrophic. Thus, it seems crucial that even basic, but potentially unboundedly self-improving[5], AIs be given at least a seed of moral values. This may be called the *moral burden assignment problem*.

Figure 4 recapitulates the different subproblems of our roadmap for the value-loading problem.

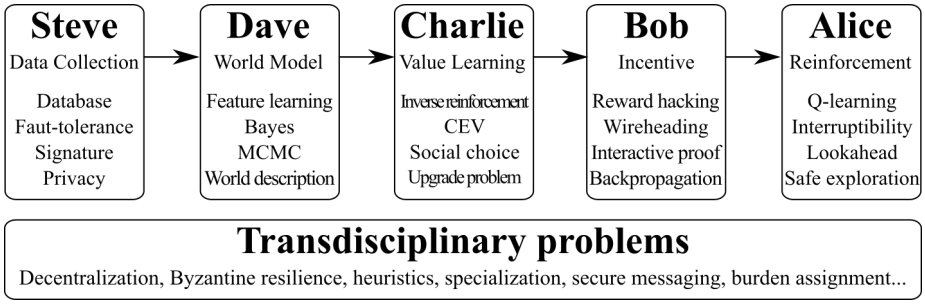

Figure 4: We propose to decompose value-loading into 5 steps. Each step is associated with further substeps or techniques. Also, there are critical subproblems that will likely be useful for several of the 5 steps.

4 Non-technical challenges

As discussed, the value-loading problem poses a large number of technical challenges that will likely require years, if not decades, of interdisciplinary collaborations. Such large-scale projects usually raise numerous non-technical challenges that will also require great manpower. It is probably worthwhile to mention them to stress the fact that it is definitely possible to greatly contribute to the global effort to solve the value-loading problem without doing AI research.

4.1 Gain respectability

Unfortunately, these days, discussions about AI safety are plagued by what Nils Nilsson dubbed the *respectability bias* [10]. Because of a long-standing poor track record in AI predictions, any reasoning about AI progress is often labeled as nonsense and dismissed as irrelevant. Worse, critics often isolate and mock the weirdest of all singularitarians' claims, as though it was representative of all discussions about human-level AIs.

Again, it should be stressed that the relevance of the value-loading problem does not require the certainty of achieving human-level AI. Even if there is only a 1% chance that AI will reach human-level within the next decade, this still is a vastly greater probability than a collision with a large asteroid. Yet, denying

[5]In particular, nonparametric AIs should perhaps be treated differently from parametric ones.

this 1% chance seems quite presumptuous, especially given the poor track record of AI predictions. Human-level AI should thus be taken extremely seriously.

Besides, even if human-level AI is not regarded as probable nor worrysome, as argued in Section 2, the value-loading problem will surely have many great applications. In its global form, it is about taking a more global view at the world to better identify the most critical actions to undertake to do good. You do not need to care about human-level AI to find this useful.

For research on value-loading to really take off, it seems extremely important that this line of work no longer be regarded as foolish nonsense. It needs to become *respectable* to publicly claim that one is working on value-loading. *Value-loading must become a mainstream field of research.*

There are many possible actions to take to make this happen. It seems that the most important action is to publicly take position in favour of research on value-loading, whether it is in papers, conferences, medias, podcasts, YouTube, Twitter or in blogs. But we need to be careful about distinguishing justifiable concerns about human-level AI from religious-sounding poorly grounded variants. It is extremely important to be pedagogical, charitable and convincing about why the possibility human-level AI must be taken seriously.

4.2 Improve debating

It seems quite clear that today's main hurdle to taking the value-loading problem seriously is that we humans are very poor at debating. Many prefer to show how ridiculous some ideas of some singularitarians are than to take the possibility of human-level AI seriously.

Part of the reason why we debate poorly is that we often replace virtuous debating by virtue signaling, especially when it comes to moral debates. We often try to stress our goodness, rather than to contribute constructively. This is a poor reflex that we all need to work on. We need to promote better debating.

Another recurrent cause of poor debating is overconfidence, which makes us inattentive to opposing sides' remarks. In fact, Johnson and Fowler write that "humans show many psychological biases, but one of the most consistent, powerful and widespread is overconfidence" [30]. It seems crucial to raise awareness about this flaw of ours, and to promote more careful thinking.

Evidently, there are many other aspects of poor debating that should be addressed, but we will not go into further details here. It should be stressed, however, that improving critical thinking and quality debating seem to be a major priority towards promoting, and thus solving, the value-loading problem.

4.3 Funding and recruiting

Given the stakes of the value-loading problem and the challenges it poses, it seems crucial that many more people get involved in the effort than is the case today. To make great progress in the value-loading problem, it is necessary to develop larger funding structures, as well as to attract talents from different areas.

To search for fundings, it seems essential to raise awareness of the importance of the value-loading problem within and beyond academic spheres. It may be worthwhile to gain influence in companies and governments. Perhaps over the years, it could even be possible to make it a social priority. In particular, it seems to be a good first step that Google decided to present a few deontological principles about its AI research [50]. Hopefully, others will follow.

Given that the research in value-loading is still in a very early phase, it is probably important as well to allow for the exploration of many alternatives — perhaps even those that do not seem promising at first. We do not yet have solid grounds to build on. We should thus not be to rigid about how to move forward.

In order to attract the best and brightest, it is probably a great idea to define as many intermediary challenging, but solvable, problems as possible. This is what has been done by [34]. It is surely a priority to define different landmark problems to nurture some competition between talents, as has been done for empirical work by ImageNet [12] and CIFAR [33]. To attract "mathematical talents", it is probably worth challenging mathematicians with non-trivial elegant mathematical conjectures related to AI safety, even when such conjectures are not guaranteed to be helpful to value-loading. Perhaps some of the numerous problems we discussed in the paper can be turned into such conjectures.

4.4 Training

Unfortunately, the value-loading problem requires a lot of background knowledge, especially in terms of machine learning concepts. Fostering a wide community of experts will require a lot of easily accessible educative materials. More universities probably should set up more lectures on AI safety. And more workshops should probably be organized.

It also seems important to reach out for manpower beyond the traditional alleys of academia. These days, massive online open courses have been able to reach out to millions of students. Many similar online resources have allowed many to learn the basics of machine learning. In particular, YouTube videos have been raising awareness and educating millions of people with surprisingly few resources, through channels like 3Blue1Brown, Siraj Raval, Computerphile, Robert Miles, ZettaBytes or Science4All.

Finally, a great way to learn and to make progress is to share open source codes and open data sets. This can also be regarded as a major step towards a massive collaboration on today's most pressing problems.

5 Conclusions

This paper discussed the *value-loading problem*, that is, the problem of encoding moral values into AIs. This is often regarded as a long-term problem. But we argued that, even if it is likely to be so, it should not be regarded as such. Indeed, it seems reasonable to assign a nonnegligible probability on the fact

that human-level AI could arise within a decade. As a result, to mitigate this near-worst case scenario, it seems urgent to make it a priority line of research as of now.

The paper also presented a general roadmap to tackle this issue. Interestingly, this roadmap identifies 5 critical steps, as well as many relevant aspects of these 5 steps. In other words, we have presented a large number of small problems that readers are highly encouraged to tackle. We hope that combining the solutions to these small problems could help to partially address the value-loading problem.

Finally, we presented non-technical challenges so that both experts and non-experts can contribute to the global effort aside from AI research. This should be of great interest to anyone who wishes to do good, including many different charities and associations.

We hope to have raised awareness of the importance of the value-loading problem and of the possible paths to partially solve it. Most importantly, we hope to have convinced you that much more work, funding and manpower is essential. And we hope that you will do your best to contribute as much as you can.

Acknowledgment. The author would like to thank El Mahdi El Mhamdi, Henrik Aslund, Sébastien Rouault and Alexandre Maurer for fruitful discussions.

References

[1] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.

[2] Sanjeev Arora, Carsten Lund, Rajeev Motwani, Madhu Sudan, and Mario Szegedy. Proof verification and the hardness of approximation problems. *Journal of the ACM (JACM)*, 45(3):501–555, 1998.

[3] Kenneth J Arrow. A difficulty in the concept of social welfare. *Journal of political economy*, 58(4):328–346, 1950.

[4] László Babai. Trading group theory for randomness. In *Proceedings of the seventeenth annual ACM symposium on Theory of computing*, pages 421–429. ACM, 1985.

[5] Leemon Baird. Hashgraph consensus: fair, fast, byzantine fault tolerance. Technical report, Swirlds Tech Report, 2016.

[6] Michel Balinski and Rida Laraki. *Majority judgment: measuring, ranking, and electing*. MIT press, 2011.

[7] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks

against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 387–402. Springer, 2013.

[8] Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. In *Advances in Neural Information Processing Systems*, pages 119–129, 2017.

[9] Paul Bloom. *Against Empathy: The Case for Rational Compassion*. Ecco, 2016.

[10] Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. OUP Oxford, 2014.

[11] Georgios Damaskinos, El Mahdi El Mhamdi, Rachid Guerraoui, Rhicheek Patra, Mahsa Taziki, et al. Asynchronous byzantine machine learning (the case of sgd). In *International Conference on Machine Learning*, pages 1153–1162, 2018.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. Ieee, 2009.

[13] David A Drachman. Do we have brain to spare? *Neurology*, 64(12):2004–2005, 2005.

[14] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[15] El Mahdi El Mhamdi, Rachid Guerraoui, Hadrien Hendrikx, and Alexandre Maurer. Dynamic safe interruptibility for decentralized multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 130–140, 2017.

[16] Owain Evans, Andreas Stuhlmüller, and Noah D Goodman. Learning the preferences of ignorant, inconsistent agents. In *AAAI*, pages 323–329, 2016.

[17] Allan Gibbard. Manipulation of voting schemes: a general result. *Econometrica: journal of the Econometric Society*, pages 587–601, 1973.

[18] Justin Gilmer, Luke Metz, Fartash Faghri, Samuel S Schoenholz, Maithra Raghu, Martin Wattenberg, and Ian Goodfellow. Adversarial spheres. *arXiv preprint arXiv:1801.02774*, 2018.

[19] Shafi Goldwasser, Silvio Micali, and Charles Rackoff. The knowledge complexity of interactive proof systems. *SIAM Journal on computing*, 18(1):186–208, 1989.

[20] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[21] Katja Grace, John Salvatier, Allan Dafoe, Baobao Zhang, and Owain Evans. When will AI exceed human performance? evidence from AI experts. *arXiv preprint arXiv:1705.08807*, 2017.

[22] Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell. The off-switch game. *arXiv preprint arXiv:1611.08219*, 2016.

[23] Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. Cooperative inverse reinforcement learning. In *Advances in neural information processing systems*, pages 3909–3917, 2016.

[24] Jonathan Haidt. *The righteous mind: Why good people are divided by politics and religion*. Vintage, 2012.

[25] Lê Nguyên Hoang. Strategy-proofness of the randomized condorcet voting system. *Social Choice and Welfare*, 48:679–701, 2017.

[26] Lê Nguyên Hoang, François Soumis, and Georges Zaccour. Measuring unfairness feeling in allocation problems. *Omega*, 65:138–147, 2016.

[27] Lê Nguyên Hoang. *La formule du savoir : une philosophie unifiée du savoir fondée sur le théorème de Bayes*. EDP Sciences, 2018.

[28] Chong Huang, Peter Kairouz, Xiao Chen, Lalitha Sankar, and Ram Rajagopal. Context-aware generative adversarial privacy. *Entropy*, 19(12):656, 2017.

[29] Neal Jean, Marshall Burke, Michael Xie, W Matthew Davis, David B Lobell, and Stefano Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794, 2016.

[30] Dominic DP Johnson and James H Fowler. The evolution of overconfidence. *Nature*, 477(7364):317, 2011.

[31] Daniel Kahneman. *Thinking, fast and slow*. Farrar, Straus and Giroux New York, 2011.

[32] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[33] Alex Krizhevsky and G Hinton. Convolutional deep belief networks on cifar-10. *Unpublished manuscript*, 40(7), 2010.

[34] Jan Leike, Miljan Martic, Victoria Krakovna, Pedro A Ortega, Tom Everitt, Andrew Lefrancq, Laurent Orseau, and Shane Legg. AI safety gridworlds. *arXiv preprint arXiv:1711.09883*, 2017.

[35] Yaniv Leviathan and Yossi Matias. Google duplex: An AI system for accomplishing real-world tasks over the phone, 2018.

[36] Cheng-Yuan Liou, Jau-Chi Huang, and Wen-Chie Yang. Modeling word perception using the elman network. *Neurocomputing*, 71(16-18):3150–3157, 2008.

[37] Hendrik P Lopuhaa, Peter J Rousseeuw, et al. Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *The Annals of Statistics*, 19(1):229–248, 1991.

[38] Daniel Lowd and Christopher Meek. Adversarial learning. In *International Conference on Machine Learning*, pages 641–647. ACM, 2005.

[39] Jarryd Martin, Tom Everitt, and Marcus Hutter. Death and suicide in universal artificial intelligence. In *Artificial General Intelligence*, pages 23–32. Springer, 2016.

[40] El Mahdi El Mhamdi, Rachid Guerraoui, and Sébastien Rouault. The hidden vulnerability of distributed learning in byzantium. In *International Conference on Machine Learning*, pages 3518–3527, 2018.

[41] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[42] Gorden E. Moore. Cramming more components onto integrated circuits. *Electronics Magazine*, 1965.

[43] Satoshi Nakamoto. Bitcoin: A peer-to-peer electronic cash system, 2008.

[44] J von Neumann and Oskar Morgenstern. *Theory of games and economic behavior*. Princeton: Princeton, 1944.

[45] Andrew Y Ng, Stuart J Russell, et al. Algorithms for inverse reinforcement learning. In *Icml*, pages 663–670, 2000.

[46] Future of Life Institute. Autonomous weapons: an open letter from AI & robotics researchers, 2015.

[47] Cathy O'Neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books, 2016.

[48] World Health Organization et al. Death and daly estimates for 2004 by cause for who member states. 2009.

[49] Laurent Orseau and MS Armstrong. Safely interruptible agents. In *Uncertainty in Artificial Intelligence: 32nd Conference (UAI 2016), edited by Alexander Ihler and Dominik Janzing*, page 557–566, 2016.

[50] Sundar Pichai. AI at Google: our principles, 2018.

[51] Rob Price. Microsoft is deleting its AI chatbot's incredibly racist tweets. *Business Insider*, 2016.

[52] Stuart Russell, Daniel Dewey, and Max Tegmark. Research priorities for robust and beneficial artificial intelligence. *AI Magazine*, 36(4):105–114, 2015.

[53] Anders Sandberg and Nick Bostrom. Global catastrophic risks survey. *Civil wars*, 98(30), 2008.

[54] Mark Allen Satterthwaite. Strategy-proofness and arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of economic theory*, 10(2):187–217, 1975.

[55] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

[56] Edward H Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 238–241, 1951.

[57] Peter Singer. *The most good you can do: How effective altruism is changing ideas about living ethically*. Text Publishing, 2015.

[58] Nate Soares. The value learning problem. In *Ethics for Artificial IntelligenceWorkshop at 25th International Joint Conference on Artificial Intelligence*, 2016.

[59] Jiawei Su, Danilo Vasconcellos Vargas, and Sakurai Kouichi. One pixel attack for fooling deep neural networks. *arXiv preprint arXiv:1710.08864*, 2017.

[60] Max Tegmark. Life 3.0. *Being Human in the Age of Artificial Intelligence. NY: Allen Lane*, 2017.

[61] Alan Turing. Computing machinery and intelligence. *Mind*, 59(236):433, 1950.

[62] Stanislaw Ulam. John von neumann 1903-1957. *Bulletin of the American mathematical society*, 64(3):1–49, 1958.

[63] Vernor Vinge. The coming technological singularity: How to survive in the post-human era. *VISION-21 Symposium*, 1993.

[64] Tobias Wängberg, Mikael Böörs, Elliot Catt, Tom Everitt, and Marcus Hutter. A game-theoretic analysis of the off-switch game. In *International Conference on Artificial General Intelligence*, pages 167–177. Springer, 2017.

[65] Robert Wiblin and Katja Grace. How well can we actually predict the future? Katja Grace on why expert opinion isn't a great guide to AI's impact and how to do better, 2018.

[66] Eliezer Yudkowsky. Coherent extrapolated volition. *Singularity Institute for Artificial Intelligence*, 2004.

[67] Eliezer Yudkowsky. Artificial intelligence as a positive and negative factor in global risk. *Global catastrophic risks*, 1(303), 2008.

[68] Barret Zoph and Quoc V. Le. Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*, 2017.