Towards Robust End-to-End Alignment

Lê Nguyên Hoang
EPFL
Chemin Alan Turing, Lausanne 1015, Switzerland

Abstract

Robust alignment is arguably both critical and extremely challenging. Loosely, it is the problem of designing algorithmic systems with strong guarantees of always being beneficial for mankind. In this paper, we propose a preliminary research program to address it in a reinforcement learning framework. This roadmap aims at decomposing the end-to-end alignment problem into numerous more tractable subproblems. We hope that each subproblem is sufficiently orthogonal to others to be tackled independently, and that combining the solutions to all such subproblems may yield a solution to alignment.

Introduction

As they are becoming more and more capable and ubiquitous, AIs are raising numerous concerns, including fairness, privacy, filter bubbles, addiction, job displacement or even existential risks (Russell, Dewey, and Tegmark 2015; Tegmark 2017). It has been argued that aligning the goals of AI systems with humans' preferences would be an efficient way to make them reliably beneficial and to avoid potential catastrophic risks (Bostrom 2014; Hoang 2018a). In fact, given the global influence of today's large-scale recommender systems (Kramer, Guillory, and Hancock 2014), it already seems urgent to propose even partial solutions to alignment.

Unfortunately, it has also been argued that alignment is an extremely difficult problem. In fact, (Bostrom 2014) argues that it "is a research challenge worthy of some of the next generation's best mathematical talent". To address it, the *Future of Life Institute* proposed a landscape of AI safety research[1]. Meanwhile, (Soares 2015; Soares and Fallenstein 2017) listed important ideas in this line of work. We hope that this paper will contribute to outline the main challenges posed by alignment.

In particular, we shall introduce a complete research program to robustly align AIs. Robustness here refers to numerous possible failure modes, including overfitting, hazardous exploration, evasion attacks, poisoning attacks, crash tolerance, Byzantine resilience, reward hacking and wireheading. To guarantee such a robustness, we argue that it is desirable to structure (at least conceptually) our AI systems in their entirety. This motivated us to propose a *roadmap* for robust end-to-end alignment.

While much of our proposal is speculative, we believe that several of the ideas presented here will be critical for 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.

Given the complexity of the problem, our roadmap here will likely be full of gaps and false good ideas. It is important to note that our purpose is not to propose a definite perfect solution. Rather, we aim at presenting a sufficiently good starting point for others to build upon.

The Roadmap

Our roadmap consists of identifying key steps to alignment. For the sake of exposition, these steps will be personified by 5 characters, called Alice, Bob, Charlie, Dave and Erin. Roughly speaking, Erin will be collecting data from the world, Dave will use these data to infer the likely states of the world, Charlie will compute the desirability of the likely 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 decomposition is graphically represented in Figure 1.

Figure 1: We decompose the alignment problem into 5 key steps: data collection, world model inference, desirability learning, incentive design and reinforcement learning.

Evidently, Alice, Bob, Charlie, Dave and Erin need not be 5 different AIs. Typically, it may be much more computationally efficient to merge Charlie and Dave. Nevertheless, at least for pedagogical reasons, it seems useful to first dissociate the different roles that these AIs have.

In the sequel, we shall further detail the challenges posed by each of the 5 AIs. We shall also argue that, for robustness and scalability reasons, these AIs will need to be further divided into many more AIs. We will see that this raises additional challenges. We shall also make a few non-technical remarks, before concluding.

[1] https://futureoflife.org/landscape/

Alice's Reinforcement learning

It seems that today's most promising framework for large-scale AIs 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 and turns it into s_{t+1}. The transition from s_t to s_{t+1} given action a_t is usually assumed to be nondeterministic. In any case, the AI then receives a reward R_{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+1}.

Note that this is a very general framework. In fact, we humans are arguably (at least partially) 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+1}, which will motivate us to update our inner model of the world θ_t into θ_{t+1}.

Let us call Alice the AI in charge of performing this reinforcement learning reasoning. Alice can thus be viewed as an algorithm, which inputs observed states s_t and rewards R_t, and undertakes actions a_t so as to typically maximize some discounted sum of expected future rewards.

Such actions will probably be mostly of the form of messages sent through the Internet. This may sound benign. But it is not. The YouTube recommender system might suggest billions of antivax videos, causing a major decrease of vaccination and an uprise of deadly diseases. Worse, if an AI is in control of 3D-printers, then a message that tells them to construct killer drones to cause a genocide would be catastrophic. On a brighter note, if an AI now promotes convincing eco-friendly messages every day to billions of people, the public opinion on climate change may greatly change.

Note that, as opposed to all other components, in some sense, Alice is the real danger. Indeed, in our framework, she is the only one that really undertakes actions. More precisely, 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 (Orseau and Armstrong 2016; El Mhamdi et al. 2017) have proposed to restrict the learning capabilities of Alice to provide provable desirable properties. Typically, they proposed to allow only a subclass of learning algorithms, i.e. of update rules of θ_{t+1} as a function of $(\theta_t, a_t, s_{t+1}, R_{t+1})$. However, such restrictions might be too costly. And this may be a big problem.

Indeed, there is already a race between competing companies in competing countries to construct powerful AIs. While it might be possible for some countries to impose some restrictions to some AIs of some companies, it is unlikely that *all* companies of all countries will accept to be restricted, especially if the restrictions are too constraining. In fact, AI safety will be useful only if the most powerful AIs are *all* subject to safety measures. As a result, the safety measures that are proposed should not be too constraining. In other words, *there are constraints on the safety constraints that can be imposed*. This is what makes AI safety so challenging.

As a result, what is perhaps more interesting are the ideas proposed by (Amodei et al. 2016) 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*.

But this is not all. Given that AIs will likely be based on machine learning, and given the lack of verification methods for AIs obtained by machine learning, we should not expect AIs to be correct all the time. Just like humans, AIs will likely be sometimes wrong. But this is extremely worrysome. Indeed, even if an AI is right 99.9999% of the time, it will still be wrong one time out of a million. Yet, AIs like recommender systems or autonomous cars take billions of decisions every day. In such cases, thousands of AI decisions may be unboundedly wrong every day!

This problem can become even more worrysome if we take into account the fact that hackers may attempt to take advantage of the AIs' deficiencies. Such hackers may typically submit only data that corresponds to cases where the AIs are wrong. This is known as *evasion attacks* (Lowd and Meek 2005; Su, Vargas, and Kouichi 2017; Gilmer et al. 2018). To avoid evasion attacks, it is crucial for an AI to never be unboundedly wrong, e.g. by reliably measuring its own confidence in its decisions and to ask for help in cases of great uncertainty.

Now, even if Alice is well-designed, she will only be an effective optimization algorithm. Unfortunately, this is no guarantee of safety or alignment. Typically, because of humans' well-known addiction to echo chambers (Haidt 2012), a watch-time maximization YouTube recommender AI may amplify filter bubbles, which may lead to worldwide geopolitical tensions. Both misaligned and unaligned AIs will likely lead to very undesirable consequences.

In fact, (Bostrom 2014) even argues that, to best reach its goals, any sufficiently strategic AI will likely first aim at so-called *instrumental goals*, e.g. gaining vastly more resources and guaranteeing self-preservation. But this is very unlikely to be in humans' best interests. In particular, it will likely motivate the AI to undertake actions that we would not regard as desirable.

To make sure that Alice will want to behave as we want her to, it seems critical to at least partially control the observed state s_{t+1} or the reward R_{t+1}. 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 undesirable, and by rewarding them when the sequence (s_t, a_t, s_{t+1}) is desirable.

Whether or not Alice's observed state s_t is constrained, her rewards R_t are clearly critical. They are her incentives, and will thus determine 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 key to alignment. Our roadmap to solve it identifies 4 key steps incarnated by Erin, Dave, Charlie and Bob.

Erin'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 undertake desirable actions, 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 the Facebook recommender system. 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 Erin's job. Typically, Erin 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 (Jean et al. 2016), 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-resilient consensus algorithms like Blockchain (Nakamoto 2008) or Hashgraph (Baird 2016), 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* (Dwork, Roth, and others 2014), or weaker versions like *generative-adversarial privacy* (Huang et al. 2017). 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 a superintelligence[2].

Dave's world model problem

Unfortunately, raw data are usually 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 Erin's collected data. 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* (Mikolov et al. 2013), *autoencoders* (Liou, Huang, and Yang 2008) and *generative adversarial networks* (GANs) (Goodfellow et al. 2014).

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

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* (Simpson 1951), and boil down to the saying "correlation is not causation". It seems crucial that 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 incentives for many actors to bias Erin's data gathering, and to thus fool Dave. This is known as *poisoning attacks* (Blanchard et al. 2017; Mhamdi, Guerraoui, and Rouault 2018; Damaskinos et al. 2018). 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 (Hoang 2018b). 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:

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

One typical instance of Dave's job is the problem of inferring global health from a wide variety of collected data. This is what has been done by (Institute for Health Metrics and Evaluation (IHME), University of Washington 2016), using a sophisticated Bayesian model that reconstructed the likely causes of deaths in countries where data were lacking.

Importantly, Bayes rule also 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 be regarded as writing a compact report of all likely states of the world, given Erin's collected data. 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.

Note also that Erin and Dave are likely to gain cognitive capabilities over time. It is surely worthwhile to anticipate the complexification of Erin's data and of Dave's

world models. It seems unclear so far how to do so. Some high-level (purely descriptive) language to describe world models is probably needed. In addition, this high-level language may need to be flexible enough 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.

Charlie's desirability learning problem

Given any of Dave's world models, Charlie's job will then be to compute how desirable this world model is. This is the *desirability learning* problem (Soares 2016), also known as *value learning*[3]. This is the problem of assigning desirability scores to different world models. These desirability scores can then serve as the basis for any agent to determine *beneficial* actions.

Unfortunately, determining what, say, the median human considers desirable is an extremely difficult problem. But again, it should be stressed that we should not aim at deriving an *ideal* inference of what people desire. This is likely to be a hopeless endeavor. Rather, we should try our best to make sure that Charlie's desirability scores will be *good enough* to avoid catastrophic outcomes, e.g. world destruction, global sufferance or major discrimination.

One proposed solution to infer human preferences is so-called *inverse reinforcement learning* (Ng, Russell, and others 2000; Evans, Stuhlmüller, and Goodman 2016). 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 humans' preferences 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 Erin'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 scores 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 desirability of different states of the world. Humans' preferences 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.

[3]To avoid raising eyebrows, we shall try to steer away from polarizing terminologies like values, moral or ethics.

This is known as a *social choice* problem. In its general form, it is the problem of aggregating the preferences of a group of disagreeing people into a single preference for the whole 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 (Arrow 1950) or the Gibbard-Satterthwaite theorem (Gibbard 1973; Satterthwaite 1975). 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 (Hoang 2017).

One particular proposal, known as *majority judgment* (Balinski and Laraki 2011), may be of particular interest to us here. Its basic idea 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' desirability scores for that state. This yields a distribution of humans' preferences for the state of the world. Majority judgment then concludes that the group's score is the quantile q of this distribution. If $q = 1/2$, this corresponds to the score 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 scoring of a state of the world will be less than a number $score$, if more than 10% of the people believe that this state should be given a score less than $score$. 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' scores. It is not clear that $score = 5$ given by John has a meaning comparable to Jane's $score = 5$. In fact, according to a theorem by von Neumann and Morgenstern (Neumann and Morgenstern 1944), 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 (Hoang, Soumis, and Zaccour 2016). 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 preferences 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 (Kahneman 2011). We tend to first have instinctive reactions to stories or facts (Bloom 2016), which quickly becomes the position we will want to defend at all costs (Haidt 2012). Worse, we are unfortunately largely unaware of why we believe or want what we believe or want. This means that our current preferences are unlikely to be what we would prefer, if we were more informed, thought more deeply, and tried to make sure our preferences were as well-founded as possible.

And arguably, we should prefer what we would prefer to prefer, rather than what we instinctively prefer. Typically, one might prefer to watch a cat video, even though one might prefer to prefer mathematics videos over cat videos. Desirablity scores should arguably encode what we would prefer

to prefer, rather than what we instinctively prefer.

To understand, 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 what he finds desirable, and analyzed all imaginable data of the world. Arguably, me^{++}'s desirability score is "more right" than *current me*'s.

This can be illustrated by the fact that past standards are often no longer regarded as desirable. Our intuitions about the desirability 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 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 are the basis of *coherent extrapolated volition* (Yudkowsky 2004). The basic idea is that we should aim at the preferences that future versions of ourselves would eventually adopt, if they were vastly more informed, had much more time to ponder what they regard as desirable, and tried their best to be better versions of themselves. In some sense, instead of making *current me*'s debate about what's desirable (which often turns 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 preferences through some social choice mechanism. This is the *preference aggregation problem*.

It is noteworthy that we clearly have epistemic uncertainty about me^{++}'s. Determining me^{++}'s desirability scores may be called the *coherent extrapoled 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 desirability scores by a probability distributions of likely desirability scores.

Such scores could also be approximated using a large number of proxies, as is done by *boosting methods* (Arora, Hazan, and Kale 2012). The use of several proxies could avoid the overfitting of any proxy. Typically, rather than relying solely on DALYs (Organization and others 2009), we probably should invoke 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). Still another approach may consist of analyzing "typical" human preferences, e.g. by using *collaborative filtering* techniques (Ricci, Rokach, and Shapira 2015). 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 intuitive preferences, so that they still basically reflect what we really regard as preferable, 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 desirable. 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 desirable things that the current versions of ourselves find repelling. Unfortunately though, we humans tend to react poorly to disagreeing jugments. And this is likely to hold even when the oppositions are our better selves. This poses a great scientific and engineering challenge. How can one be best convinced of the judgments 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 improvement problem*.

To address this issue, (Irving, Christiano, and Amodei 2018) have discussed the possibility of setting up a debate between opposing AIs. In particular, they asked whether a human judge would be able to lean towards the better AI for the right reasons. Interestingly, such a debate might allow for significantly more powerful "proofs of superiority" than monologues, at least if the analogy with the so-called *polynomial hierarchy* of complexity theory holds.

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

Bob's incentive design

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 Erin, Dave and Charlie. Evidently he could simply compute the expectation of Charlie's scores for the likely states of the world. But this is probably a bad idea, as it opens the door to *reward hacking*.

Recall 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 Erin, 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 Erin's data collection, one way for Alice to increase her rewards would be to feed Erin 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 Erin, Dave and Charlie. This could potentially be extremely bad for mankind. Indeed, especially if Alice cares about discounted future rewards, she might eventually regard mankind as a possible threat to her objective.

This is why it is of the utmost importance that Alice's incentives be (partially) aligned with Erin, 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 correlate with Charlie's scores, they also give Alice the incentives to guarantee that Erin, 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 Erin, Dave and Charlie for the better. Ideally, she would even want them to be computationally more powerful than herself, especially in the long run. This approach would bear resemblance with the idea of *self-nudge* (Thaler and Sunstein 2009). This corresponds to strategies that we humans sometimes use to nudge ourselves (or others) into doing what we want to want to do, rather than what our latest emotion or laziness invites us to do.

Unfortunately, it seems unclear how Bob can best 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. Typically, Bob should make sure Alice's rewards are affected by the reliability of Erin's data. The more reliable Erin's data, the larger Alice's rewards. Similarly, when Dave or Charlie feel that their computations 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 scores. 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 proofs* (IP) (Babai 1985; Goldwasser, Micali, and Rackoff 1989) or *probabilistically checkable proofs* (PCP) (Arora et al. 1998) might be useful to force Alice to prove its correct behavior. By requesting such proofs to yield large rewards, Bob might be able to incentivize Alice's transparency. All such considerations make up Bob's *incentive problem*.

It may or may not be useful to enable Bob to switch off Alice. It should be stressed though that (safe) interruptibility is nontrivial, as discussed by (Orseau and Armstrong 2016; El Mhamdi et al. 2017; Martin, Everitt, and Hutter 2016; Hadfield-Menell et al. 2016a; 2016b; Wängberg et al. 2017) 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 preferable 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 score or Erin's reliability. Having said this, we leave open the technical question of how to best design this.

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

Decentralization

We have decomposed alignment into 5 components for the sake of exposition. However, any component will likely have to be decentralized to gain reliability and scalability. In other words, instead of having a single Alice, a single Bob, a single Charlie, a single Dave and a single Erin, it seems crucial to construct multiple Alices, Bobs, Charlies, Daves and Erins.

This is key to *crash-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 remain mostly the same. But crash-tolerance is likely to be insufficient. Instead, we should design *Byzantine-resilient* mechanisms, that is, mechanisms that still perform correctly despite the presence of hacked or malicious Bobs. Estimators with large statistical breakdowns (Lopuhaa, Rousseeuw, and others 1991), e.g. (geometric) medians and variants (Blanchard et al. 2017), may be useful for this purpose.

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. More generally, the careful design of secure communication channels between the components of the AIs seems key. This may be called the *secure messaging problem*.

Another difficulty is the addition of more powerful and precise Bobs, Charlies, Daves and Erins to the pipeline. It is not yet clear how to best integrate reliable new comers, especially given that such new comers may be malicious. In fact, they may want to first act benevolent 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*, which was recently discussed by (Christiano, Shlegeris, and Amodei 2018) who proposed using numerous weaker AIs to supervise stronger AIs. More research in this direction is probably needed.

Now, in addition to reliability, decentralization may also enable different Alices, Bobs, Charlies, Daves and Erins to focus on specific tasks. This would allow to separate different problems, which could lead to more optimized solutions at lower costs. 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.

Another open question is the extent to which AIs should be exposed to Bobs' rewards. Typically, if a small company creates its own AI, to what extent should this AI be aligned? It should be noted that this may be computationally very costly, as it may be hard to separate the signal of interest to the AI from the noise of Bobs' rewards. Intuitively, the more influential an 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 to avoid any *diffusion of responsibility*, i.e. many small AIs that disregard safety concerns on the ground that they each hardly have any

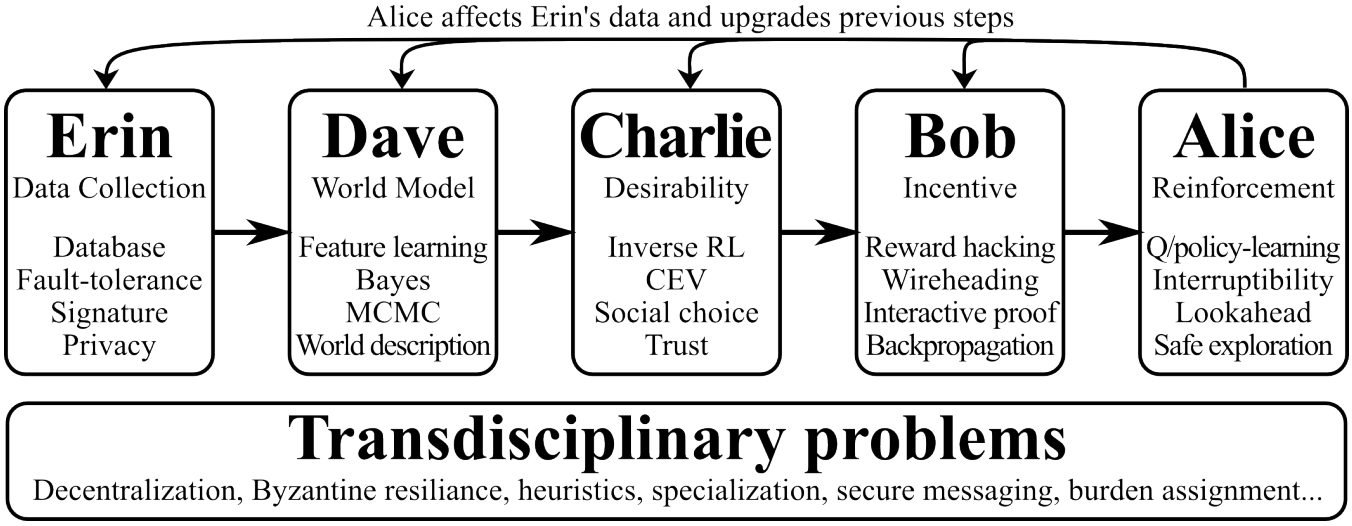

Figure 2: We propose to decompose alignment into 5 steps. Each step is associated with further substeps or techniques. Also, there are critical subproblems that will likely be useful for several of the 5 steps.

global impact on the world.

What makes this nontrivial is that any AI may gain capability and influence over time. An unaligned weak AI could eventually become an unaligned human-level AI. To avoid this, even basic, but potentially unboundedly self-improving[5] AIs should be given at least a seed of alignment, which may grow as AIs become more powerful. More generally, AIs should strike a balance between some original (possibly unaligned) objective and the importance they give to alignment. This may be called the *alignment burden assignment problem*.

Figure 2 recapitulates our complete roadmap.

Non-technical challenges

Given the difficulty of alignment, its resolution will surely require solving a large number of non-technical challenges as well. We briefly mention some of them here.

Perhaps most important is the lack of respectability that is sometimes associated with this line of research. For alignment to be solved, it needs to gain respectability from the scientific community, and perhaps beyond this community as well. This is why it seems to be of the utmost importance that discussions around alignment be carried out carefully to avoid confusions.

Evidently, alignment definitely needs much more manpower, which will require funding and recruiting. It seems particularly important to attract mathematical talents towards this line of work. This evidently also raises the challenge of training as many brilliant minds as possible.

Finally, questions around AI, AI safety and moral philosophy are sadly often poorly debated. There often is a lot of overconfidence, and a lack of well-founded reasoning. For alignment research to gain momentum, it seems crucial to make debating more informative, respectful and stimulating.

Conclusion

This paper discussed the *alignment problem*, that is, the problem of aligning the goals of AIs with human preferences. It 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 hopefully more tractable subproblems that readers are highly encouraged to tackle. We hope that combining the solutions to these subproblems could help to partially address alignment. And we hope that any reader will be able to better determine how he or she may best contribute to the global effort[6].

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

References

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

Arora, S.; Lund, C.; Motwani, R.; Sudan, M.; and Szegedy, M. 1998. Proof verification and the hardness of approximation problems. *Journal of the ACM (JACM)* 45(3):501–555.

Arora, S.; Hazan, E.; and Kale, S. 2012. The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing* 8(1):121–164.

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

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

[6] Please note that a more complete version of this paper is also available (Hoang 2018b).

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

Baird, L. 2016. Hashgraph consensus: fair, fast, byzantine fault tolerance. Technical report, Swirlds Tech Report.

Balinski, M., and Laraki, R. 2011. *Majority judgment: measuring, ranking, and electing*. MIT press.

Blanchard, P.; El Mhamdi, E. M.; Guerraoui, R.; and Stainer, J. 2017. Machine learning with adversaries: Byzantine tolerant gradient descent. In *Advances in Neural Information Processing Systems*, 119–129.

Bloom, P. 2016. *Against Empathy: The Case for Rational Compassion*. Ecco.

Bostrom, N. 2014. *Superintelligence: Paths, Dangers, Strategies*. OUP Oxford.

Christiano, P.; Shlegeris, B.; and Amodei, D. 2018. Supervising strong learners by amplifying weak experts. In review.

Damaskinos, G.; El Mhamdi, E. M.; Guerraoui, R.; Patra, R.; Taziki, M.; et al. 2018. Asynchronous byzantine machine learning (the case of sgd). In *International Conference on Machine Learning*, 1153–1162.

Dwork, C.; Roth, A.; et al. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science* 9(3–4):211–407.

El Mhamdi, E. M.; Guerraoui, R.; Hendrikx, H.; and Maurer, A. 2017. Dynamic safe interruptibility for decentralized multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, 130–140.

Evans, O.; Stuhlmüller, A.; and Goodman, N. D. 2016. Learning the preferences of ignorant, inconsistent agents. In *AAAI*, 323–329.

Gibbard, A. 1973. Manipulation of voting schemes: a general result. *Econometrica: journal of the Econometric Society* 587–601.

Gilmer, J.; Metz, L.; Faghri, F.; Schoenholz, S. S.; Raghu, M.; Wattenberg, M.; and Goodfellow, I. 2018. Adversarial spheres. *arXiv preprint arXiv:1801.02774*.

Goldwasser, S.; Micali, S.; and Rackoff, C. 1989. The knowledge complexity of interactive proof systems. *SIAM Journal on computing* 18(1):186–208.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680.

Hadfield-Menell, D.; Dragan, A.; Abbeel, P.; and Russell, S. 2016a. The off-switch game. *arXiv preprint arXiv:1611.08219*.

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

Haidt, J. 2012. *The righteous mind: Why good people are divided by politics and religion*. Vintage.

Hoang, L. N.; Soumis, F.; and Zaccour, G. 2016. Measuring unfairness feeling in allocation problems. *Omega* 65:138–147.

Hoang, L. N. 2017. Strategy-proofness of the randomized condorcet voting system. *Social Choice and Welfare* 48:679–701.

Hoang, L. N. 2018a. A roadmap for the value-loading problem. *arXiv preprint arXiv:1809.01036*.

Hoang, L. N. 2018b. *La formule du savoir : une philosophie unifiée du savoir fondée sur le théorème de Bayes*. EDP Sciences. English translation forthcoming.

Huang, C.; Kairouz, P.; Chen, X.; Sankar, L.; and Rajagopal, R. 2017. Context-aware generative adversarial privacy. *Entropy* 19(12):656.

Institute for Health Metrics and Evaluation (IHME), University of Washington. 2016. Gbd compare data visualization.

Irving, G.; Christiano, P.; and Amodei, D. 2018. Ai safety via debate. *arXiv preprint arXiv:1805.00899*.

Jean, N.; Burke, M.; Xie, M.; Davis, W. M.; Lobell, D. B.; and Ermon, S. 2016. Combining satellite imagery and machine learning to predict poverty. *Science* 353(6301):790–794.

Kahneman, D. 2011. *Thinking, fast and slow*. Farrar, Straus and Giroux New York.

Kramer, A. D.; Guillory, J. E.; and Hancock, J. T. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences* 201320040.

Liou, C.-Y.; Huang, J.-C.; and Yang, W.-C. 2008. Modeling word perception using the elman network. *Neurocomputing* 71(16-18):3150–3157.

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

Lowd, D., and Meek, C. 2005. Adversarial learning. In *International Conference on Machine Learning*, 641–647. ACM.

Martin, J.; Everitt, T.; and Hutter, M. 2016. Death and suicide in universal artificial intelligence. In *Artificial General Intelligence*. Springer. 23–32.

Mhamdi, E. M. E.; Guerraoui, R.; and Rouault, S. 2018. The hidden vulnerability of distributed learning in byzantium. In *International Conference on Machine Learning*, 3518–3527.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Nakamoto, S. 2008. Bitcoin: A peer-to-peer electronic cash system.

Neumann, J. v., and Morgenstern, O. 1944. *Theory of games and economic behavior*. Princeton: Princeton.

Ng, A. Y.; Russell, S. J.; et al. 2000. Algorithms for inverse reinforcement learning. In *Icml*, 663–670.

Organization, W. H., et al. 2009. Death and daly estimates for 2004 by cause for who member states.

Orseau, L., and Armstrong, M. 2016. Safely interruptible agents. In *Uncertainty in Artificial Intelligence: 32nd Conference (UAI 2016), edited by Alexander Ihler and Dominik Janzing*, 557–566.

Ricci, F.; Rokach, L.; and Shapira, B. 2015. Recommender systems: introduction and challenges. In *Recommender systems handbook*. Springer. 1–34.

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

Satterthwaite, M. A. 1975. 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.

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

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

Soares, N. 2015. Aligning superintelligence with human interests: An annotated bibliography. *Intelligence* 17(4):391–444.

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

Su, J.; Vargas, D. V.; and Kouichi, S. 2017. One pixel attack for fooling deep neural networks. *arXiv preprint arXiv:1710.08864*.

Tegmark, M. 2017. Life 3.0. *Being Human in the Age of Artificial Intelligence. NY: Allen Lane*.

Thaler, R., and Sunstein, C. 2009. *Nudge: Improving Decisions About Health, Wealth, and Happiness*. Penguin Books.

Wängberg, T.; Böörs, M.; Catt, E.; Everitt, T.; and Hutter, M. 2017. A game-theoretic analysis of the off-switch game. In *International Conference on Artificial General Intelligence*, 167–177. Springer.

Yudkowsky, E. 2004. Coherent extrapolated volition. *Singularity Institute for Artificial Intelligence*.