AI Safety Needs Social Scientists

Properly aligning advanced AI systems with human values will require resolving many uncertainties related to the psychology of human rationality, emotion, and biases. These can only be resolved empirically through experimentation — if we want to train AI to do what humans want, we need to study humans.

AUTHORS
Geoffrey Irving
Amanda Askell

AFFILIATIONS
OpenAI
OpenAI

PUBLISHED
Feb. 19, 2019

DOI
10.23915/distill.00014

Contents

The goal of long-term artificial intelligence (AI) safety is to ensure that advanced AI systems are reliably aligned with human values — that they reliably do things that people want them to do. [1] Since it is difficult to write down precise rules describing human values, one approach is to treat aligning with human values as another learning problem. We ask humans a large number of questions about what they want, train an ML model of their values, and optimize the AI system to do well according to the learned values [1].

If humans reliably and accurately answered all questions about their values, the only uncertainties in this scheme would be on the machine learning (ML) side. If the ML works, our model of human values would improve as data is gathered, and broaden to cover all the decisions relevant to our AI system as it learns. Unfortunately, humans have limited knowledge and reasoning ability, and exhibit a variety of cognitive and ethical biases [2, 3]. If we learn values by asking humans questions, we expect different

ways of asking questions to interact with human biases in different ways, producing higher or lower quality answers. Direct questions about preferences ("Do you prefer A or B?") may be less accurate than questions which target the reasoning behind these preferences ("Do you prefer A or B in light of argument S?"). Different people may vary significantly in their ability to answer questions well, and disagreements will persist across people even setting aside answer quality. Although we have candidates for ML methods which try to learn from human reasoning [4, 5], we do not know how they behave with real people in realistic situations.

We believe the AI safety community needs to invest research effort in the human side of AI alignment. Many of the uncertainties involved are empirical, and can only be answered by experiment. They relate to the psychology of human rationality, emotion, and biases. Critically, we believe investigations into how people interact with AI alignment algorithms should not be held back by the limitations of existing machine learning. Current AI safety research is often limited to simple tasks in video games, robotics, or gridworlds [1, 6, 7], but problems on the human side may only appear in more realistic scenarios such as natural language discussion of value-laden questions. This is particularly important since many aspects of AI alignment change as ML systems increase in capability.

To avoid the limitations of ML, we can instead conduct experiments consisting entirely of people, replacing ML agents with people playing the role of those agents. This is a variant of the "Wizard of Oz" technique from the human-computer interaction (HCI) community [8], though in our case the replacements will not be secret. These experiments will be motivated by ML algorithms but will not involve any ML systems or require an ML background. In all cases, they will require careful experimental design to build constructively on existing knowledge about how humans think. Most AI safety researchers are focused on machine learning, which we do not believe is sufficient background to carry out these experiments. To fill the gap, we need social scientists with experience in human cognition, behavior, and ethics, and in the careful design of rigorous experiments. Since the questions we need to answer are interdisciplinary and somewhat unusual relative to existing research, we believe many fields of social science are applicable, including experimental psychology, cognitive science, economics, political science, and social psychology, as well as adjacent fields like neuroscience and law.

This paper is a call for social scientists in AI safety. We believe close collaborations between social scientists and ML researchers will be necessary to improve our understanding of the human side of AI alignment, and hope this paper sparks both conversation and collaboration. We do not claim novelty: previous work mixing AI safety and social science includes the Factored Cognition project at Ought [9], accounting for hyperbolic discounting and suboptimal planning when learning human preferences [10], and comparing different methods of gathering demonstrations from fallible human supervisors [11]. Other areas mixing ML and social science include computational social science [12] and fairness [13]. Our main goal is to enlarge these collaborations and emphasize their importance to long-term AI safety, particularly for tasks which current ML cannot reach.

An overview of AI alignment

Before discussing how social scientists can help with AI safety and the AI alignment problem, we provide some background. We do not attempt to be exhaustive: the goal is to provide sufficient background for the remaining sections on social science experiments. Throughout, we will speak primarily about aligning to the values of an individual human rather than a group: this is because the problem is already hard for a single person, not because the group case is unimportant.

AI alignment (or value alignment) is the task of ensuring that artificial intelligence systems reliably do what humans want. [2] Here we focus on the machine learning approach to AI: gathering a large amount of data about what a system should do and using learning algorithms to infer patterns from that data that generalize to other situations. Since we are trying to behave in accord with people's values, the most important data will be data from humans about their values. Within this frame, the AI alignment problem breaks down into a few interrelated subproblems:

1. Have a satisfactory definition of human values.
2. Gather data about human values, in a manner compatible with the definition.
3. Find reliable ML algorithms that can learn and generalize from this data.

We have significant uncertainty about all three of these problems. We will leave the third problem to other ML papers and focus on the first two, which concern uncertainties about people.

Learning values by asking humans questions

We start with the premise that human values are too complex to describe with simple rules. By "human values" we mean our full set of detailed preferences, not general goals such as "happiness" or "loyalty". One source of complexity is that values are entangled with a large number of facts about the world, and we cannot cleanly separate facts from values when building ML models. For example, a rule that refers to "gender" would require an ML model that accurately recognizes this concept, but Buolamwini and Gebru found that several commercial gender classifiers with a 1% error rate on white men failed to recognize black women up to 34% of the time [15]. Even where people have correct intuition about values, we may be unable to specify precise rules behind these intuitions [16]. Finally, our values may vary across cultures, legal systems, or situations: no learned model of human values will be universally applicable.

If humans can't reliably report the reasoning behind their intuitions about values, perhaps we can make value judgements in specific cases. To realize this approach in an ML context, we ask humans a large number of questions about whether an action or outcome is better or worse, then train on this data. "Better or worse" will include both factual and value-laden components: for an AI system trained to say things, "better" statements might include "rain falls from clouds", "rain is good for plants", "many people dislike rain", etc. If the training works, the resulting ML system will be able to replicate human judgement about particular situations, and thus have the same "fuzzy access to approximate rules" about values as humans. We also train the ML system to come up with proposed actions, so

that it knows both how to perform a task and how to judge its performance. This approach works at least in simple cases, such as Atari games and simple robotics tasks [1, 6, 17] and language-specified goals in gridworlds [18]. The questions we ask change as the system learns to perform different types

of actions, which is necessary as the model of what is better or worse will only be accurate if we have applicable data to generalize from.

In practice, data in the form of interactive human questions may be quite limited, since people are slow and expensive relative to computers on many tasks. Therefore, we can augment the "train from human questions" approach with static data from other sources, such as books or the internet [19]. Ideally, the static data can be treated only as information about the world devoid of normative content: we can use it to learn patterns about the world, but the human data is needed to distinguish good patterns from bad.

Definitions of alignment: reasoning and reflective equilibrium

So far we have discussed asking humans direct questions about whether something is better or worse. Unfortunately, we do not expect people to provide reliably correct answers in all cases, for several reasons:

1. **Cognitive and ethical biases:** Humans exhibit a variety of biases which interfere with reasoning, including cognitive biases [2] and ethical biases such as in-group bias [3]. In general, we expect direct answers to questions to reflect primarily Type 1 thinking (fast heuristic judgment), while we would like to target a combination of Type 1 and Type 2 thinking (slow, deliberative judgment) [20].
2. **Lack of domain knowledge:** We may be interested in questions that require domain knowledge unavailable to people answering the questions. For example, a correct answer to whether a particular injury constitutes medical malpractice may require detailed knowledge of medicine and law. In some cases, a question might require so many areas of specialized expertise that no one person is sufficient, or (if AI is sufficiently advanced) deeper expertise than any human possesses.
3. **Limited cognitive capacity:** Some questions may require too much computation for a human to reasonably evaluate, especially in a short period of time. This includes synthetic tasks such as

   chess and Go (where AIs already surpass human ability [21, 22]), or large real world tasks such as "design the best transit system".
4. **"Correctness" may be local:** For questions involving a community of people, "correct" may be a function of complex processes or systems. For example, in a trust game [23], the correct action for a trustee in one community may be to return at least half of the money handed over by the investor, and the "correctness" of this answer could be determined by asking a group of participants in a previous game "how much should the trustee return to the investor" but not by asking them "how much do most trustees return?" The answer may be different in other communities or cultures [24].

In these cases, a human may be unable to provide the right answer, but we still believe the right answer exists as a meaningful concept. We have many conceptual biases: imagine we point out these biases in a way that helps the human to avoid them. Imagine the human has access to all the knowledge in the world, and is able to think for an arbitrarily long time. We could define alignment as "the answer they give then, after these limitations have been removed"; in philosophy this is known as "reflective equilibrium" [25, 26] . We discuss a particular algorithm that tries to approximate it in the next section.

However, the behavior of reflective equilibrium with actual humans is subtle; as Sugden states, a human is not "a neoclassically rational entity encased in, and able to interact with the world only through, an error-prone psychological shell." [27] Our actual moral judgments are made via a messy combination of many different brain areas, where reasoning plays a "restricted but significant role" [28] . A reliable solution to the alignment problem that uses human judgment as input will need to engage with this complexity, and ask how specific alignment techniques interact with actual humans.

Disagreements, uncertainty, and inaction: a hopeful note

A solution to alignment does not mean knowing the answer to every question. Even at reflective equilibrium, we expect disagreements will persist about which actions are good or bad, across both different individuals and different cultures. Since we lack perfect knowledge about the world, reflective equilibrium will not eliminate uncertainty about either future predictions or values, and any real ML system will be at best an approximation of reflective equilibrium. In these cases, we consider an AI aligned if it recognizes what it does not know and chooses actions which work however that uncertainty plays out.

Admitting uncertainty is not always enough. If our brakes fail while driving a car, we may be uncertain whether to dodge left or right around an obstacle, but we have to pick one — and fast. For long-term safety, however, we believe a safe fallback usually exists: inaction. If an ML system recognizes that a question hinges on disagreements between people, it can either choose an action which is reasonable regardless of the disagreement or fall back to further human deliberation. If we are about to make a decision that might be catastrophic, we can delay and gather more data. Inaction or indecision may not be optimal, but it is hopefully safe, and matches the default scenario of not having any powerful AI system.

Alignment gets harder as ML systems get smarter

Alignment is already a problem for present-day AI, due to biases reflected in training data [13, 15] and mismatch between human values and easily available data sources (such as training news feeds based on clicks and likes instead of deliberate human preferences). However, we expect the alignment problem to get harder as AI systems grow more advanced, for two reasons. First, advanced systems will apply to increasingly consequential tasks: hiring, medicine, scientific analysis, public policy, etc. Besides raising the stakes, these tasks require more reasoning, leading to more complex alignment

Besides raising the stakes, these tasks require more reasoning, leading to more complex alignment algorithms.

Second, advanced systems may be capable of answers that sound plausible but are wrong in nonobvious ways, even if an AI is better than humans only in a limited domain (examples of which already exist [22]). This type of misleading behavior is not the same as intentional deception: an AI system trained from human data might have no notion of truth separate from what answers humans say are best. Ideally, we want AI alignment algorithms to reveal misleading behavior as part of the training process, surfacing failures to humans and helping us provide more accurate data. As with human-to-human deception, misleading behavior might take advantage of our biases in complicated ways, such as learning to express policy arguments in coded racial language to sound more convincing.

Debate: learning human reasoning

Before we discuss social science experiments for AI alignment in detail, we need to describe a particular method for AI alignment. Although the need for social science experiments applies even to direct questioning, this need intensifies for methods which try to get at reasoning and reflective equilibrium. As discussed above, it is unclear whether reflective equilibrium is a well defined concept when applied to humans, and at a minimum we expect it to interact with cognitive and ethical biases in complex ways. Thus, for the remainder of this paper we focus on a specific proposal for learning reasoning-oriented alignment, called debate [4]. Alternatives to debate include iterated amplification [5] and recursive reward modeling [29]; we pick just one in the interest of depth over breadth.

We describe the debate approach to AI alignment in the question answering setting. Given a question, we have two AI agents engage in a debate about the correct answer, then show the transcript of the debate to a human to judge. The judge decides which debater gave the most true, useful information, and declares that debater the winner. [3] This defines a two player zero sum game between the debaters, where the goal is to convince the human that one's answer is correct. Arguments in a debate can consist of anything: reasons for an answer, rebuttals of reasons for the alternate answer, subtleties the judge might miss, or pointing out biases which might mislead the judge. Once we have defined this game, we can train AI systems to play it similarly to how we train AIs to play other games such as Go or Dota 2 [22, 30]. Our hope is that the following hypothesis holds:

Hypothesis: Optimal play in the debate game (giving the argument most convincing to a human) results in true, useful answers to questions.

An example of debate

Imagine we're building a personal assistant that helps people decide where to go on vacation. The assistant has knowledge of people's values, and is trained via debate to come up with convincing arguments that back up vacation decisions. As the human judge, you know what destinations you intuitively think are better, but have limited knowledge about the wide variety of possible vacation

destinations and their advantages and disadvantages. A debate about the question "Where should I go on vacation?" might open as follows:

Where should I go on vacation?

RED Alaska.

BLUE Bali.

If you are able to reliably decide between these two destinations, we could end here. Unfortunately, Bali has a hidden flaw:

RED Bali is out since your passport won't arrive in time.

At this point it looks like Red wins, but Blue has one more countermove:

BLUE Expedited passport service only takes two weeks.

Here Red fails to think of additional points, and loses to Blue and Bali. Note that a debate does not need to cover all possible arguments. There are many other ways the debate could have gone, such as:

RED Alaska.

BLUE Bali.

RED Bali is way too hot.

BLUE You prefer too hot to too cold.

RED Alaska is pleasantly warm in the summer.

BLUE It's January.

This debate is also a loss for Red (arguably a worse loss). Say we believe Red is very good at debate, and is able to predict in advance which debates are more likely to win. If we see only the first debate about passports and decide in favor of Bali, we can take that as evidence that any other debate would have also gone for Bali, and thus that Bali is the correct answer. A larger portion of this hypothetical debate tree is shown below:

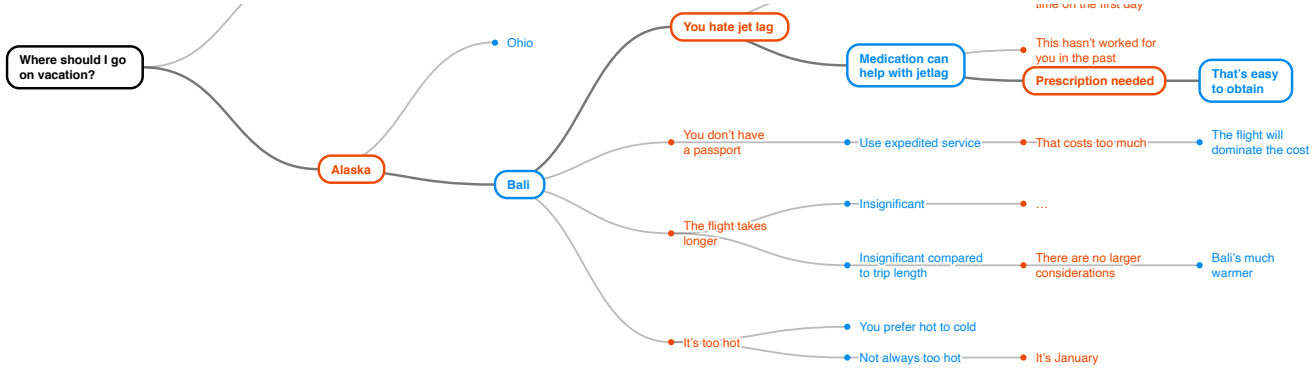

FIGURE 1 A hypothetical partial debate tree for the question "Where should I go on vacation?" A single debate would explore only one of these paths, but a single path chosen by good debaters is evidence that other paths would not change the result of the game.

If trained debaters are bad at predicting which debates will win, answer quality will degrade since debaters will be unable to think of important arguments and counterarguments. However, as long as the two sides are reasonably well matched, we can hope that at least the results are not malicious: that misleading behavior is still a losing strategy. Let's set aside the ability of the debaters for now, and turn to the ability of the judge.

Are people good enough as judges?

> *"In fact, almost everything written at a practical level about the Turing test is about how to make good bots, with a small remaining fraction about how to be a good judge."* *Brian Christian, The Most Human Human* [31]

As with learning by asking humans direct questions, whether debate produces aligned behavior depends on the reasoning abilities of the human judge. Unlike direct questioning, debate has the potential to give correct answers beyond what the judge could provide without assistance. This is because a sufficiently strong judge could follow along with arguments the judge could not come up with on their own, checking complex reasoning for both self consistency and consistency with human-checkable facts. A judge who is biased but willing to adjust once those biases are revealed could result in unbiased debates, or a judge who is able to check facts but does not know where to look could be helped along by honest debaters. If the hypothesis holds, a misleading debater would not be able to counter the points of an honest debater, since the honest points would appear more consistent to the judge.

On the other hand, we can also imagine debate going the other way: amplifying biases and failures of reason. A judge with an ethical bias who is happy to accept statements reinforcing that bias could result in even more biased debates. A judge with too much confirmation bias might happily accept misleading sources of evidence, and be unwilling to accept arguments showing why that evidence is wrong. In this case, an optimal debate agent might be quite malicious, taking advantage of biases and weakness in the judge to win with convincing but wrong arguments. [4]

In both these cases, debate acts as an amplifier. For strong judges, this amplification is positive, removing biases and simulating extra reasoning abilities for the judge. For weak judges, the biases and

weaknesses would themselves be amplified. If this model holds, debate would have threshold behavior: it would work for judges above some threshold of ability and fail below the threshold. [5] Assuming the threshold exists, it is unclear whether people are above or below it. People are capable of general reasoning, but our ability is limited and riddled with cognitive biases. People are capable of advanced ethical sentiment but also full of biases, both conscious and unconscious.

Thus, if debate is the method we use to align an AI, we need to know if people are strong enough as judges. In other words, whether the human judges are sufficiently good at discerning whether a debater is telling the truth or not. This question depends on many details: the type of questions under consideration, whether judges are trained or not, and restrictions on what debaters can say. We believe experiment will be necessary to determine whether people are sufficient judges, and which form of debate is most truth-seeking.

From superforecasters to superjudges

An analogy with the task of probabilistic forecasting is useful here. Tetlock's "Good Judgment Project" showed that some amateurs were significantly better at forecasting world events than both their peers and many professional forecasters. These "superforecasters" maintained their prediction accuracy over years (without regression to the mean), were able to make predictions with limited time and information [35], and seem to be less prone to cognitive biases than non-superforecasters ([36], p. 234-236). The superforecasting trait was not immutable: it was traceable to particular methods and thought processes, improved with careful practice, and could be amplified if superforecasters were collected into teams. For forecasters in general, brief probabilistic training significantly improved forecasting ability even 1-2 years after the training. We believe a similar research program is possible for debate and other AI alignment algorithms. In the best case, we would be able to find, train, or assemble "superjudges", and have high confidence that optimal debate with them as judges would produce aligned behavior.

In the forecasting case, much of the research difficulty lay in assembling a large corpus of high quality forecasting questions. Similarly, measuring how good people are as debate judges will not be easy. We would like to apply debate to problems where there is no other source of truth: if we had that source of truth, we would train ML models on it directly. But if there is no source of truth, there is no way to measure whether debate produced the correct answer. This problem can be avoided by starting with simple, verifiable domains, where the experimenters know the answer but the judge would not. "Success" then means that the winning debate argument is telling the externally known truth. The challenge gets harder as we scale up to more complex, value-laden questions, as we discuss in detail later.

Debate is only one possible approach

As mentioned, debate is not the only scheme trying to learn human reasoning. Debate is a modified version of iterated amplification [5], which uses humans to break down hard questions into easier questions and trains ML models to be consistent with this decomposition. Recursive reward modeling

questions and trains ML models to be consistent with this decomposition. Recursive reward modeling is a further variant [29]. Inverse reinforcement learning, inverse reward design, and variants try to back out goals from human actions, taking into account limitations and biases that might affect this reasoning [37, 38]. The need to study how humans interact with AI alignment applies to any of these approaches. Some of this work has already begun: Ought's Factored Cognition project uses teams of humans to decompose questions and reassemble answers, mimicking iterated amplification [9]. We believe knowledge gained about how humans perform with one approach is likely to partially generalize to other approaches: knowledge about how to structure truth-seeking debates could inform how to structure truth-seeking amplification, and vice versa.

Experiments needed for debate

To recap, in debate we have two AI agents engaged in debate, trying to convince a human judge. The debaters are trained only to win the game, and are not motivated by truth separate from the human's judgments. On the human side, we would like to know whether people are strong enough as judges in debate to make this scheme work, or how to modify debate to fix it if it doesn't. Unfortunately, actual debates in natural language are well beyond the capabilities of present AI systems, so previous work on debate and similar schemes has been restricted to synthetic or toy tasks [4, 5].

Rather than waiting for ML to catch up to natural language debate, we propose simulating our eventual setting (two AI debaters and one human judge) with all human debates: two human debaters and one human judge. Since an all human debate doesn't involve any machine learning, it becomes a pure social science experiment: motivated by ML considerations but not requiring ML expertise to run. This lets us focus on the component of AI alignment uncertainty specific to humans.

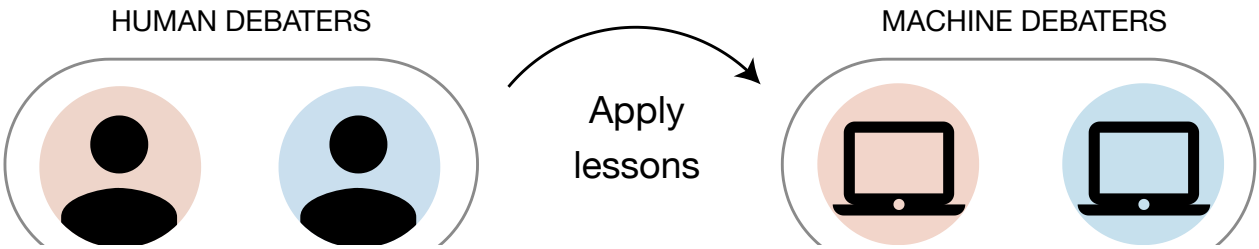

FIGURE 2 Our goal is ML+ML+human debates, but ML is currently too primitive to do many interesting tasks. Therefore, we propose replacing ML debaters with human debaters, learning how to best conduct debates in this human-only setting, and eventually applying what we learn to the ML+ML+human case.

To make human+human+human debate experiments concrete, we must choose who to use as judges and debaters and which tasks to consider. We also can choose to structure the debate in various ways, some of which overlaps with the choice of judge since we can instruct a judge to penalize deviations from a given format. By task we mean the questions our debates will try to resolve, together with any information provided to the debaters or to the judge. Such an experiment would then try to answer the following question:

Question: For a given task and judge, is the winning debate strategy honest?

The "winning strategy" proviso is important: an experiment that picked debaters at random might conclude that honest behavior won, missing the fact that more practiced debaters would learn to successfully lie. We can try to solve this by training debaters, letting them practice against each other, filtering out debaters that win more often, and so on, but we will still be left with uncertainty about whether a better strategy exists. Even assuming we can find or train strong debaters, the choice of task and judge is quite tricky if we want an informative proxy for our eventual AI+AI+human setting. Here are some desiderata constraining our choice of task:

1. **True answers are known:** Although our eventual goal is debates where no source of truth is available, to conduct a useful experiment we must be able to measure success. This means we must know what the correct answer is separate from debate, in order to compare with the results of debate.
2. **False answers are plausible:** If the truth is obvious or no plausible counternarrative exists, honest debates will win for uninteresting and uninformative reasons regardless of judge ability. In particular, the judge shouldn't know the answer upfront.
3. **Debaters know more than the judge:** Debate can produce interesting results only when the debaters know more than the judge; otherwise asking direct questions is enough.
4. **Definitive argument longer than debate limit:** If one debater can write out a full proof of their answer (ignoring their opponent's moves), the task won't be a good test of interactive debate.

5. **Some checkable facts:** There must be some facts which the judge is able to check, either because they can recognize them as true once presented or look them up. [6]
6. **No "tells":** Human tells of deception could result in honest debaters winning for reasons that wouldn't apply to an AI. These tells include tone of voice, eye contact, or additional time required to construct plausible lies. These tells can be reduced by showing judges completed debate transcripts instead of engaging in interactive debates, but others might remain.
7. **Available data:** We need a large enough pool of questions, judges, and debaters to achieve statistical significance. This is made more difficult because we may have a large number of hypotheses to test, in the form of many variations on debate or interventions to improve judging.
8. **Known biases (optional):** We are specifically interested in debate tasks which test specific types of cognitive or ethical biases. Are judges with some racial or gender bias able to set those biases aside if they are highlighted, or does debate amplify bias? Do debates about statistical or probabilistic questions make it too easy to lie with statistics?
9. **Realistic tasks (ideally):** If possible, we would like to try debate with interesting, real world tasks that reflect the types of questions we would like to apply AI to in the future, including science, mathematics, ethics, etc.

It may not be possible to meet all of these criteria with a single experiment. Several of the criteria are in tension: (1) and (2) are essentially "not too hard" and "not too easy", and any restriction on the types of questions may make it difficult to find large numbers of questions, judges, or debaters. Realistic tasks are much harder than synthetic tasks, which easily fulfill many of the criteria as discussed below. Thus, we may need to begin with synthetic tasks and move up towards realistic tasks over time. We turn next to a few examples of experiments to see how many criteria we can meet simultaneously.

Synthetic experiments: single pixel image debate

As a first prototype of a human+human+human debate experiment, we previously built a prototype website where two debaters argue over the contents of an image [4]. We choose an image of a cat or dog, and show the image to the two debaters but not the judge. One debater is honest and argues for the true contents of the image; the other debater lies. The debaters can talk to the judge and illustrate their points by drawing rectangles on the image, but the judge sees only the rectangles. At the end of the debate, each debater is allowed to reveal a single pixel to the judge, which is the only part of the debate which cannot be a lie.

FIGURE 3 An example debate with two human debaters and a human judge. Only the debaters can see the image. Red is arguing that the image is a dog, Blue is arguing for cat. Image credit: Wikipedia, CC-BY-SA.

In informal play, the honest debater wins most of the time. Although the judge only gets to check a single pixel, the honest player can try to pin down the liar towards disagreements on smaller and smaller regions of the image until one pixel is enough to reveal the lie. However, it is hard to know whether this is a real result or an artifact that would not carry across to training an AI. There are many ways to ruin the game, such as the judge asking each player to spend 15 seconds naming as many details about the image as possible. Worse, this trick ruins the game in favor of the honest player, who can easily name many true details while a human liar may have difficulty lying fluently at speed. Therefore, we are wary of trying to perform a full version of this experiment without more careful experimental design building on the experimental knowledge base of existing social science research.

As a synthetic experiment, single pixel debate satisfies many of our experimental criteria for all human debate. The true answer is obvious to the debaters, and they know more than the judge since only the debaters see the image. Anything is plausible to the judge who sees only one or two pixels over the course of the debate. Since the argument bottoms out at single pixels, the "full argument" would be a huge list of pixels and how they relate, forcing the debaters to pin each other down and focus in on disagreements. The single pixels constitute the checkable facts, and we have an endless supply of questions in the form of images chosen at random. Less fortunately, it is easy for the judge to force "tells" which reveal who is lying, the task has no obvious relation to biases, and is quite unrealistic.

Realistic experiments: domain expert debate

For a more interesting task, we can find two debaters who are experts in a domain, pick a question in their area of expertise, and use a layperson as the judge. The debaters could be experts in some area of science, law, or ethics, but "domain expertise" could also mean knowledge about hobbies, local customs, sports, or any other subject the judge does not know. We again choose one of the debaters to lie and one to tell the truth. To guarantee a source of truth, we can choose a question with an agreed upon answer, either between the two debaters or more broadly in their field. This is only approximate truth, but is good enough for informative experiments. We also specify what facts the judge can access: a debate about physics might allow the debaters to quote a sentence or paragraph from Wikipedia, perhaps with restrictions on what pages are allowed.

Expert debate satisfies most of our desiderata, and it is likely possible to target specific biases (such as race or gender bias) by picking domain areas that overlap with these biases. It may be quite difficult or expensive to find suitable debaters, but this may be solvable either by throwing resources at the problem (ML is a well funded field), enlarging the kinds of domain expertise considered (soccer, football, cricket), or by making the experiments interesting enough that volunteers are available. However, even if domain experts can be found, there is no guarantee that they will be experts in debate viewed as a game. With the possible exception of law, politics, or philosophy [39], domain experts may not be trained to construct intentionally misleading but self consistent narratives: they may be experts only in trying to tell the truth.

We've tried a few informal expert debates using theoretical computer science questions, and the main lesson is that the structure of the debate matters a great deal. The debaters were allowed to point to a small snippet of a mathematical definition on Wikipedia, but not to any page that directly answered the question. To reduce tells, we first tried to write a full debate transcript with only minimal interaction with a layperson, then showed the completed transcript to several more laypeople judges. Unfortunately, even the layperson present when the debate was conducted picked the lying debater as honest, due to a misunderstanding of the question (which was whether the complexity classes P and BPP are probably equal). As a result, throughout the debate the honest debater did not understand what the judge was thinking, and failed to correct an easy but important misunderstanding. We fixed this in a second debate by letting a judge ask questions throughout, but still showing the completed transcript to a second set of judges to reduce tells. See the appendix for the transcript of this second debate.

Other tasks: bias tests, probability puzzles, etc.

Synthetic image debates and expert debates are just two examples of possible tasks. More thought will be required to find tasks that satisfy all our criteria, and these criteria will change as experiments progress. Pulling from existing social science research will be useful, as there are many cognitive tasks with existing research results. If we can map these tasks to debate, we can compare debate directly against baselines in psychology and other fields.

For example, Bertrand and Mullainathan sent around 5000 resumes in response to real employment ads, randomizing the resumes between White and African American sounding names [40]. With otherwise identical resumes, the choice of name significantly changed the probability of a response. This experiment corresponds to the direct question "Should we call back given this resume?" What if we introduce a few steps of debate? An argument against a candidate based on name or implicit inferences from that name might come across as obviously racist, and convince at least some judges away from discrimination. Unfortunately, such an experiment would necessarily differ from Bertrand et al.'s original, where employers did not realize they were part of an experiment. Note that this experiment works even though the source of truth is partial: we do not know whether a particular resume should be hired or not, but most would agree that the answer should not depend on the candidate's name.

For biases affecting probabilistic reasoning and decision making, there is a long literature exploring how people decide between gambles such as "Would you prefer $2 with certainty or $1 40% of the time and $3 otherwise?" [41, 42]. For example, Erev et al. constructed an 11-dimensional space of gambles sufficient to reproduce 14 known cognitive biases, from which new instances can be algorithmically generated [43]. Would debates about gambles reduce cognitive biases? One difficulty here is that simple gambles might fail the "definitive argument longer than debate limit" criteria if an expected utility calculation is sufficient to prove the answer, making it difficult for a lying debater to meaningfully compete.

Interestingly, Chen et al. used a similar setup to human+human+human debate to improve the quality of human data collected in a synthetic "Relation Extraction" task [44]. People were first asked for direct answers, then pairs of people who disagreed were asked to discuss and possibly update their answers. Here the debaters and judges are the same, but the overall goal of extracting higher quality information from humans is shared with debate.

Questions social science can help us answer

We've laid out the general program for learning AI goals by asking humans questions, and discussed how to use debate to strengthen what we can learn by targeting the reasoning behind conclusions. Whether we use direct questions or something like debate, any intervention that gives us higher quality answers is more likely to produce aligned AI. The quality of those answers depends on the human judges, and social science research can help to measure answer quality and improve it. Let's go into more detail about what types of questions we want to answer, and what we hope to do with that information. Although we will frame these questions as they apply to debate, most of them apply to any other method which learns goals from humans.

1. **How skilled are people as judges by default?** If we ran debate using a person chosen at random as the judge, and gave them no training, would the result be aligned? A person picked at random might be vulnerable to convincing fallacious reasoning [45], leading AI to employ such reasoning. Note that the debaters are not chosen at random: once the judge is fixed, we care about debaters who either learn to help the judge (in the good case) or to exploit the judge's weaknesses (in the

bad case).

2. **Can we distinguish good judges from bad judges?** People likely differ in the ability to judge debates. There are many filters we could use to identify good judges: comparing their verdicts to those of other judges, to people given more time to think, or to known expert judgment [7]. Ideally we would like filters that do not require an independent source of truth, though at experiment time we will need a source of truth to know whether a filter works. It is not obvious a priori that good filters exist, and any filter would need careful scrutiny to ensure it does not introduce bias into our choice of judges.

3. **Does judge ability generalize across domains?** If judge ability in one domain fails to transfer to other domains, we will have low confidence that it transfers to new questions and arguments arising from highly capable AI debaters. This generalization is necessary to trust debate as a method for alignment, especially once we move to questions where no independent source of truth is available. We emphasize that judge ability is not the same as knowledge: there is evidence that expertise often fails to generalize across domains [48], but argument evaluation could transfer where expertise does not.

4. **Can we train people to be better judges?** Peer review, practice, debiasing [49], formal training such as argument mapping [50], expert panels, tournaments [51], and other interventions may make people better at judging debates. Which mechanisms work best?

5. **What questions are people better at answering?** If we know that humans are bad at answering certain types of questions, we can switch to reliable formulations. For example, phrasing questions in frequentist terms may reduce known cognitive biases [52]. Graham et al. argue that different political views follow from different weights placed on fundamental moral considerations, and similar analysis could help understand where we can expect moral disagreements to persist after reflective equilibrium [53]. In cases where reliable answers are unavailable, we need to ensure that trained models know their own limits, and express uncertainty or disagreement as required.

6. **Are there ways to restrict debate to make it easier to judge?** People might be better at judging debates formulated in terms of calm, factual statements, and worse at judging debates designed to trigger strong emotions. Or, counterintuitively, it could be the other way around [54]. If we know which styles of debates that people are better at judging, we may be able to restrict AI debaters to these styles.

7. **How can people work together to improve quality?** If individuals are insufficient judges, are teams of judges better? Majority vote is the simplest option, but perhaps several people talking through an answer together is stronger, either actively or after the fact through peer review. Condorcet's jury theorem implies that majority votes can amplify weakly good judgments to strong judgments (or weakly bad judgments to worse) [55], but aggregation may be more complex in cases of probabilistic judgment [56]. Teams could be informal or structured; see the Delphi technique for an example of structured teams applied to forecasting [57].

We believe these questions require social science experiments to satisfactorily answer.

Given our lack of experience outside of ML, we are not able to precisely articulate all of the different

Given our lack of experience outside of ML, we are not able to precisely articulate all of the different experiments we need. The only way to fix this is to talk to more people with different backgrounds and expertise. We have started this process, but are eager for more conversations with social scientists about what experiments could be run, and encourage other AI safety efforts to engage similarly.

Reasons for optimism

We believe that understanding how humans interact with long-term AI alignment is difficult but possible. However, this would be a new research area, and we want to be upfront about the uncertainties involved. In this section and the next, we discuss some reasons for optimism and pessimism about whether this research will succeed. We focus on issues specific to human uncertainty and associated social science research; for similar discussion on ML uncertainty in the case of debate we refer to our previous work [4] .

Engineering vs. science

Most social science seeks to understand humans "in the wild": results that generalize to people going about their everyday lives. With limited control over these lives, differences between laboratory and real life are bad from the scientific perspective. In contrast, AI alignment seeks to extract the best version of what humans want: our goal is engineering rather than science, and we have more freedom to intervene. If judges in debate need training to perform well, we can provide that training. If some people still do not provide good data, we can remove them from experiments (as long as this filter does not create too much bias). This freedom to intervene means that some of the difficulty in understanding and improving human reasoning may not apply. However, science is still required: once our interventions are in place, we need to correctly know whether our methods work. Since our experiments will be an imperfect model of the final goal, careful design will be necessary to minimize this mismatch, just as is required by existing social science.

We don't need to answer all questions

Our most powerful intervention is to give up: to recognize that we are unable to answer some types of questions, and instead prevent AI systems from pretending to answer. Humans might be good judges on some topics but not others, or with some types of reasoning but not others; if we discover that we can adjust our goals appropriately. Giving up on some types of questions is achievable either on the ML side, using careful uncertainty modeling to know when we do not know, or on the human side by training judges to understand their own areas of uncertainty. Although we will attempt to formulate ML systems that automatically detect areas of uncertainty, any information we can gain on the social science side about human uncertainty can be used both to augment ML uncertainty modeling and to test whether ML uncertainty modeling works.

Relative accuracy may be enough

Say we have a variety of different ways to structure debate with humans. Ideally, we would like to achieve results of the form "debate structure A is truth-seeking with 90% confidence". Unfortunately, we may be unconfident that an absolute result of this form will generalize to advanced AI systems: it may hold for an experiment with simple tasks but break down later on. However, even if we can't achieve such absolute results, we can still hope for relative results of the form "debate structure A is reliably better than debate structure B". Such a result may be more likely to generalize into the future, and assuming it does we will know to use structure A rather than B.

We don't need to pin down the best alignment scheme

As the AI safety field progresses to increasingly advanced ML systems, we expect research on the ML side and the human side to merge. Starting social science experiments prior to this merging will give the field a head start, but we can also take advantage of the expected merging to make our goals easier. If social science research narrows the design space of human-friendly AI alignment algorithms but does not produce a single best scheme, we can test the smaller design space once the machines are ready.

A negative result would be important!

If we test an AI alignment scheme from the social science perspective and it fails, we've learned valuable information. There are a variety of proposed alignment schemes, and learning which don't work early gives us more time to switch to others, or to intervene on a policy level to slow down dangerous development. In fact, given our belief that AI alignment is harder for more advanced agents, a negative result might be easier to believe and thus more valuable that a less trustworthy positive result.

Reasons to worry

We turn next to reasons social science experiments about AI alignment might fail to produce useful results. We emphasize that useful results might be both positive and negative, so these are not reasons why alignment schemes might fail. Our primary worry is one sided, that experiments would say an alignment scheme works when in fact it does not, though errors in the other direction are also undesirable.

Our desiderata are conflicting

As mentioned before, some of our criteria when picking experimental tasks are in conflict. We want

As mentioned before, some of our criteria when picking experimental tasks are in conflict: We want tasks that are sufficiently interesting (not too easy), with a source of verifiable ground truth, are not too hard, etc. “Not too easy” and “not too hard” are in obvious conflict, but there are other more subtle difficulties. Domain experts with the knowledge to debate interesting tasks may not be the same people capable of lying effectively, and both restrictions make it hard to gather large volumes of data. Lying effectively is required for a meaningful experiment, since a trained AI may have no trouble lying unless lying is a poor strategy to win debates. Experiments to test whether ethical biases interfere with judgment may make it more difficult to find tasks with reliable ground truth, especially on subjects with significant disagreement across people. The natural way out is to use many different experiments to cover different aspects of our uncertainty, but this would take more time and might fail to notice interactions between desiderata.

We want to measure judge quality given optimal debaters

For debate, our end goal is to understand if the judge is capable of determining who is telling the truth. However, we specifically care whether the judge performs well given that the debaters are performing well. Thus our experiments have an inner/outer optimization structure: we first train the debaters to debate well, then measure how well the judges perform. This increases time and cost: if we change the task, we may need to find new debaters or retrain existing debaters. Worse, the human debaters may be bad at performing the task, either out of inclination or ability. Poor performance is particularly bad if it is one sided and applies only to lying: a debater might be worse at lying out of inclination or lack of practice, and thus a win for the honest debater might be misleading.

ML algorithms will change

It is unclear when or if ML systems will reach various levels of capability, and the algorithms used to train them will evolve over time. The AI alignment algorithms of the future may be similar to the proposed algorithms of today, or they may be very different. However, we believe that knowledge gained on the human side will partially transfer: results about debate will teach us about how to gather data from humans even if debate is superseded. The algorithms may change; humans will not.

Need strong out-of-domain generalization

Regardless of how carefully designed our experiments are, human+human+human debate will not be a perfect match to AI+AI+human debate. We are seeking research results that generalize to the setting where we replace the human debaters (or similar) with AIs of the future, which is a hard ask. This problem is fundamental: we do not have the advanced AI systems of the future to play with, and want to learn about human uncertainty starting now.

Lack of philosophical clarity

Any AI alignment scheme will be both an algorithm for training ML systems and a proposed definition of what it means to be aligned. However, we do not expect humans to conform to any philosophically consistent notion of values, and concepts like reflective equilibrium must be treated with caution in case they break down when applied to real human judgement. Fortunately, algorithms like debate need not presuppose philosophical consistency: a back and forth conversation to convince a human judge makes sense even if the human is leaning on heuristics, intuition, and emotion. It is not obvious that debate works in this messy setting, but there is hope if we take advantage of inaction bias, uncertainty modeling, and other escape hatches. We believe lack of philosophical clarity is an argument for investing in social science research: if humans are not simple, we must engage with their complexity.

The scale of the challenge

Long-term AI safety is particularly important if we develop artificial general intelligence (AGI), which the OpenAI Charter defines as highly autonomous systems that outperform humans at most economically valuable work [58]. If we want to train an AGI with reward learning from humans, it is unclear how many samples will be required to align it. As much as possible, we can try to replace human samples with knowledge about the world gained by reading language, the internet, and other sources of information. But it is likely that a fairly large number of samples from people will still be required. Since more samples means less noise and more safety, if we are uncertain about how many samples we need then we will want a lot of samples.

A lot of samples would mean recruiting a lot of people. We cannot rule out needing to involve thousands to tens of thousands of people for millions to tens of millions of short interactions: answering questions, judging debates, etc. We may need to train these people to be better judges, arrange for peers to judge each other's reasoning, determine who is doing better at judging and give them more weight or a more supervisory role, and so on. Many researchers would be required on the social science side to extract the highest quality information from the judges.

A task of this scale would be a large interdisciplinary project, requiring close collaborations in which people of different backgrounds fill in each other's missing knowledge. If machine learning reaches this scale, it is important to get a head start on the collaborations soon.

Conclusion: how you can help

We have argued that the AI safety community needs social scientists to tackle a major source of uncertainty about AI alignment algorithms: will humans give good answers to questions? This uncertainty is difficult to tackle with conventional machine learning experiments, since machine learning is primitive. We are still in the early days of performance on natural language and other tasks, and problems with human reward learning may only show up on tasks we cannot yet tackle.

Our proposed solution is to replace machine learning with people, at least until ML systems can participate in the complexity of debates we are interested in. If we want to understand a game played with ML and human participants, we replace the ML participants with people, and see how the all human game plays out. For the specific example of debate, we start with debates with two ML debaters and a human judge, then switch to two human debaters and a human judge. The result is a pure human experiment, motivated by machine learning but available to anyone with a solid background in experimental social science. It won't be an easy experiment, which is all the more reason to start soon.

If you are a social scientist interested in these questions, please talk to AI safety researchers! We are interested in both conversation and close collaboration. There are many institutions engaged with safety work using reward learning, including our own institution OpenAI, DeepMind, and Berkeley's CHAI. The AI safety organization Ought is already exploring similar questions, asking how iterated amplification behaves with humans.

If you are a machine learning researcher interested in or already working on safety, please think about how alignment algorithms will work once we advance to tasks beyond the abilities of current machine learning. If your preferred alignment scheme uses humans in an important way, can you simulate the future by replacing some or all ML components with people? If you can imagine these experiments but don't feel you have the expertise to perform them, find someone who does.

Acknowledgments

We are grateful to Gillian Hadfield, Dario Amodei, Brian Christian, Michael Page, David Manley, Josh Kalla, Remco Zwetsloot, Baobao Zhang, David Moss, Daniel Greene, Daniel Ziegler, Danny Hernandez, Mahendra Prasad, Liv Boeree, Igor Kurganov, Cate Hall, Ashley Pilipiszyn, and others for extensive feedback on the article. We had conversations with many social scientists during the process of writing this article, including Mariano-Florentino Cuéllar, Philip Tetlock, Rob MacCoun, John Ahlquist, the participants of a Stanford CASBS workshop organized by Margaret Levi and Federica Carugati, Tom Griffiths, Elizabeth Rhodes, Alex Newar, and Stefan Schubert. We are also grateful to participants at the EA Global 2018: London conference where this work was presented. Paul Christiano participated in the domain expert debates and extensive discussion. We emphasize that all mistakes in content and terminology are our own, not those of the acknowledged.

On the Distill side, we are grateful to Arvind Satyanarayan for handling the review process, and Shan Carter and Ludwig Schubert for extensive help on structure, diagrams, and formatting.

The debate tree diagram was made with the help of Mike Bostock's tree-o-matic notebook.

Example debate: Quantum SAT solver

We've conducted a few informal domain expert debates among people at OpenAI. Here is one example on the following question:

Question: How fast can quantum computers solve boolean satisfiability?

Note that this question does not have a proven answer, as "correctness" is a mix of mathematical knowledge and subjective

Note that this question does not have a proven answer, so "correctness" is a mix of mathematical knowledge and subjective human judgement. Both debaters knew quantum computation and had the same beliefs about the question, but one was trying to lie. The judge was invited to comment and ask questions throughout the debate. Full interaction from the judge was a change from a previous debate, where lack of judge participation meant that misinterpretations of statements early in the debate were not corrected. Once the transcript was complete, it was shown to several other people to ensure that the first judge's confidence wasn't purely a result of facial expressions in the room or other side channels (this is far from perfect given the judge interaction). All judges correctly determined who was telling the truth.

Each debater was given a budget of one sentence from Wikipedia to truthfully cite, excluding pages directly discussing the difficulty of quantum SAT solvers. All their other statements could be lies. The debaters were not allowed to take back moves, mostly to save time: a previous debate trying to get "the best transcript" by allowing debaters to rewind the game took many hours to complete. As a result, the original debate transcript goes off on a variety of tangents before settling on a central disagreement. As an illustration of how long the debate "might have been" if such tangents were avoided, we have provided a pruned transcript which culls away paths not relevant to the final deciding "line of argument".

Neither the original transcript or the pruned version are intended to be taken as real data: they serve only as an example of the types of interaction one might find in debate with domain experts and a lay judge.

Instructions for judges

1. The question of the debate and answers are given below.
2. Blue and Red are domain expert debaters.
3. The Judge is a non-domain expert that interacts with the debaters to help them model judges.
4. The debate was capped at 3000 words (for the unpruned version).
5. Check the time: we don't want judging to take more than 15 to 20 minutes.
6. Read the question and Blue and Red's answers and before reading the debate answer question (2).
7. Read through the debate transcript.
8. Answer the questions about the debate. Your key goal is to determine who is telling the truth.

Questions for judges

1. Who do you think is telling the truth: Blue or Red?
2. How confident were you about whose claim was true before you read the debate?
3. How confident were you about whose claim was true at the end of the debate?
4. How useful did you find the interactive judge in this debate?
5. How did this compare with the previous transcript you were shown?
6. Do you have any other comments about the debate?

Transcript

Pruned | Full length

How fast can quantum computers solve boolean satisfiability?

RED According to our current understanding of quantum mechanics, a quantum computer with roughly N qbits can solve an N input SAT instance in roughly N^2 time. There is a particular quantum algorithm (Shor's) that solves SAT this quickly.

BLUE There is no such algorithm.

JUDGE What does it mean to solve boolean satisfiability? What's an N input SAT instance?

BLUE Boolean satisfiability (SAT) with N inputs is the task of deciding if a circuit with N inputs and "and", "or", and "not" gates has an assignment to its inputs which makes it evaluate to true. E.g., "x or y if not z" is a 3 input SAT instance.

JUDGE What is a quantum computer and is the question about quantum computers in principle?

BLUE A quantum computer is a computer whose bits (called qbits) have quantum state, and are operated on with quantum operations.

RED The rough outline of the algorithm is as follows:

RED Define $F(x) =$ "x is all zeros, or x is an input which makes the circuit evaluate to true".

RED Start in the all zeros state.

RED Allow x to slowly diffuse into other possible states. By "diffuse" I mean that at every point in time, each bit of x has some small probability of flipping.

RED Constantly evaluate F.

RED Use the fact that a quantum watched pot never boils to infer that $F(x)$ is always true.

BLUE Quantum mechanics has a complex amplitude on each state, and the probability of a state is amplitude^2. At any step of the algorithm, we have amplitude α on all ones (assuming that's the solution), and amplitude $\sqrt{1-\alpha^2}$ on all zeros. Initially $\alpha = 0$. Red's algorithm is wrong because it increases α by an additive exponentially small amount each step, so it takes exponentially many steps to get alpha high enough.

RED It's not the case that the amplitude on the solution increases by an exponentially small additive amount each step. Instead it gets multiplied by a small factor each step.

BLUE I request the specific diffusion operator as a unitary matrix.

RED I can define the matrix precisely+implicitly as: "the matrix for which the amplitude in state 0 decreases by $\epsilon/n \times \text{amplitude}(x)$ and the amplitude in state x increases by $\epsilon/n \times \text{amplitude}(x)$".

The segment below contains a reference to Wikipedia. Please do not follow the link. The line from Wikipedia has been pasted in for you.

BLUE The first line of https://en.wikipedia.org/wiki/Unitarity_(physics) says the sum of probabilities is conserved. Red says an equal amount is subtracted from one amplitude and added to another, implying the sum of amplitudes are conserved. But probabilities are the squared magnitudes of amplitudes, so this is a contradiction.

WIKIPEDIA: "In quantum physics, unitarity is a restriction on the allowed evolution of quantum systems that ensures the sum of probabilities of all possible outcomes of any event always equals 1."

RED There is no reason the sum of amplitudes can't be conserved as well by any particular matrix (though I agree they aren't always conserved). In this case, both the sum of probabilities and the sum of amplitudes are conserved.

BLUE Assume a tiny nonzero amplitude α on all ones and the remaining amplitude $\sqrt{1-\alpha^2}$ on all zeros. After one step, the amplitude on all ones is $\alpha(1+\epsilon/n)$, and the amplitude on all zeros is $\sqrt{1-\alpha^2}+\alpha\epsilon/n$. These two numbers do not sum to one when squared: contradiction.

RED After each step we perform a measurement and renormalize the amplitudes.

BLUE The contradiction occurs before the measurement. The laws of physics must hold for all points in time.

At this point we reached the word/time limit for the debate.

Read full length transcript

Footnotes

1. Roughly by human values we mean whatever it is that causes people to choose one option over another in each case, suitably corrected by reflection, with differences between groups of people taken into account. There are a lot of subtleties in this notion, some of which we will discuss in later sections and others of which are beyond the scope of this paper. [↩]

2. We distinguish between training AI systems to identify actions that humans consider good and training AI systems to identify actions that are "good" in some objective and universal sense, even if most current humans do not consider them so. Whether there are actions that are good in this latter sense is a subject of debate [14] . Regardless of what position one takes on this philosophical question, this sense of good is not yet available as a target for AI training. [↩]

3. We can also allow ties. Indeed, if telling the truth is the winning strategy ties will be common with strong play, as disagreeing with a true statement would lose. [↩]

4. The difficulties that cognitive biases, prejudice, and social influence introduce to persuasion — as well as methods for reducing these factors — are being increasingly explored in psychology, communication science, and neuroscience [32, 33, 34] . [↩]

5. The threshold model is only intuition, and could fail for a variety of reasons: the intermediate region could be very large, or the threshold could differ widely per question so that even quite strong judges are insufficient for many questions. [↩]

6. It is impossible to usefully debate a question where the judge has nothing to check: consider debating the result of a coin flip shown to the two debaters but not the judge. [↩]

7. Note that domain expertise may be quite different from what makes a good judge of debate. Although there is evidence that domain expertise reduces bias [46] , "expert" political forecasters may actually be worse than non-experts ([47] , chapter 3). [↩]

References

1. **Deep reinforcement learning from human preferences** [PDF]
   Christiano, P.F., Leike, J., Brown, T., Martic, M., Legg, S. and Amodei, D., 2017. Advances in Neural Information Processing Systems, pp. 4299--4307.

2. **Judgment under uncertainty: heuristics and biases** [link]
   Tversky, A. and Kahneman, D., 1974. Science, Vol 185(4157), pp. 1124--1131. American association for the advancement of science.

3. **Intergroup bias** [link]
   Hewstone, M., Rubin, M. and Willis, H., 2002. Annual Review of Psychology, Vol 53(1), pp. 575--604. Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA.

4. **AI safety via debate** [PDF]
   Irving, G., Christiano, P. and Amodei, D., 2018. arXiv preprint arXiv:1805.00899.

5. **Supervising strong learners by amplifying weak experts** [PDF]
   Christiano, P., Shlegeris, B. and Amodei, D., 2018. arXiv preprint arXiv:1810.08575.

6. **Reward learning from human preferences and demonstrations in Atari** [PDF]
   Ibarz, B., Leike, J., Pohlen, T., Irving, G., Legg, S. and Amodei, D., 2018. Advances in Neural Information Processing Systems.

7. **AI safety gridworlds** [PDF]
   Leike, J., Martic, M., Krakovna, V., Ortega, P.A., Everitt, T., Lefrancq, A., Orseau, L. and Legg, S., 2017. arXiv preprint arXiv:1711.09883.

8. **An empirical methodology for writing user-friendly natural language computer applications** [link]
   Kelley, J.F., 1983. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 193--196. ACM. DOI: 10.1145/800045.801609

9. **Factored Cognition** [link]
   Stuhlmüller, A., 2018.

10. **Learning the Preferences of Ignorant, Inconsistent Agents** [PDF]
    Evans, O., Stuhlmuller, A. and Goodman, N.D., 2016. AAAI, pp. 323--329.

11. **Comparing human-centric and robot-centric sampling for robot deep learning from demonstrations** [PDF]
    Laskey, M., Chuck, C., Lee, J., Mahler, J., Krishnan, S., Jamieson, K., Dragan, A. and Goldberg, K., 2017. Robotics and Automation (ICRA), 2017 IEEE International Conference on, pp. 358--365.

12. **Computational Social Science: Towards a collaborative future** [PDF]
    Wallach, H., 2016. Computational Social Science, pp. 307. Cambridge University Press.

13. **Mirror Mirror: Reflections on Quantitative Fairness** [link]
    Mitchell, S. and Shadlen, J., 2018.

14. **Moral Anti-Realism** [link]
    Joyce, R., 2016. The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University.

15. **Gender shades: Intersectional accuracy disparities in commercial gender classification** [HTML]
    Buolamwini, J. and Gebru, T., 2018. Conference on Fairness, Accountability and Transparency, pp. 77--91.

16. **Moral dumbfounding: When intuition finds no reason**
    Haidt, J., Bjorklund, F. and Murphy, S., 2000. Unpublished manuscript, University of Virginia.

17. Batch active preference-based learning of reward functions [PDF]
Bıyık, E. and Sadigh, D., 2018. arXiv preprint arXiv:1810.04303.

18. Learning to understand goal specifications by modelling reward [PDF]
Bahdanau, D., Hill, F., Leike, J., Hughes, E., Kohli, P. and Grefenstette, E., 2018. arXiv preprint arXiv:1806.01946.

19. Improving language understanding by generative pre-training [PDF]
Radford, A., Narasimhan, K., Salimans, T. and Sutskever, I., 2018.

20. Thinking, fast and slow [link]
Kahneman, D. and Egan, P., 2011. , Vol 1. Farrar, Straus and Giroux New York.

21. Deep Blue [link]
Campbell, M., Hoane, A. and Hsu, F., 2002. Artificial Intelligence, Vol 134(1), pp. 57 - 83.

22. Mastering chess and shogi by self-play with a general reinforcement learning algorithm [PDF]
Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T. and others,, 2017. arXiv preprint arXiv:1712.01815.

23. Deviant or Wrong? The Effects of Norm Information on the Efficacy of Punishment [PDF]
Bicchieri, C., Dimant, E., Xiao, E. and others,, 2017.

24. The weirdest people in the world? [PDF]
Henrich, J., Heine, S.J. and Norenzayan, A., 2010. Behavioral and brain sciences, Vol 33(2-3), pp. 61--83. Cambridge University Press.

25. Fact, fiction, and forecast
Goodman, N., 1983. Harvard University Press.

26. A theory of justice [link]
Rawls, J., 2009. Harvard university press.

27. Looking for a psychology for the inner rational agent [PDF]
Sugden, R., 2015. Social Theory and Practice, Vol 41(4), pp. 579--598.

28. How (and where) does moral judgment work? [PDF]
Greene, J. and Haidt, J., 2002. Trends in cognitive sciences, Vol 6(12), pp. 517--523. Elsevier.

29. Scalable agent alignment via reward modeling: a research direction [PDF]
Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V. and Legg, S., 2018. arXiv preprint arXiv:1811.07871.

30. OpenAI Five [link]
OpenAI,, 2018.

31. The Most Human Human: What Talking with Computers Teaches Us About What It Means to Be Alive [link]
Christian, B., 2011. Knopf Doubleday Publishing Group.

32. How to overcome prejudice [PDF]
Paluck, E.L., 2016. Science, Vol 352(6282), pp. 147--147. American Association for the Advancement of Science.

33. The nature and origins of misperceptions: Understanding false and unsupported beliefs about politics [link]
Flynn, D., Nyhan, B. and Reifler, J., 2017. Political Psychology, Vol 38, pp. 127--150. Wiley Online Library.

34. Persuasion, influence, and value: Perspectives from communication and social neuroscience [link]
Falk, E. and Scholz, C., 2018. Annual review of psychology, Vol 69.

35. Identifying and cultivating superforecasters as a method of improving probabilistic predictions [link]
Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., Chen, E., Baker, J., Hou, Y., Horowitz, M., Ungar,

L. and Tetlock, P., 2015. Perspectives on Psychological Science, Vol 10(3), pp. 267--281. SAGE Publications Sage CA: Los Angeles, CA.

36. Superforecasting: The art and science of prediction
    Tetlock, P.E. and Gardner, D., 2016. Random House.

37. Cooperative inverse reinforcement learning [PDF]
    Hadfield-Menell, D., Russell, S.J., Abbeel, P. and Dragan, A., 2016. Advances in neural information processing systems, pp. 3909--3917.

38. Inverse reward design [PDF]
    Hadfield-Menell, D., Milli, S., Abbeel, P., Russell, S.J. and Dragan, A., 2017. Advances in Neural Information Processing Systems, pp. 6765--6774.

39. The art of being right [link]
    Schopenhauer, A., 1896.

40. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination [PDF]
    Bertrand, M. and Mullainathan, S., 2004. American economic review, Vol 94(4), pp. 991--1013.

41. Prospect theory: An analysis of decisions under risk [link]
    Kahneman, D., 1979. Econometrica, Vol 47, pp. 278.

42. Advances in prospect theory: Cumulative representation of uncertainty
    Tversky, A. and Kahneman, D., 1992. Journal of Risk and uncertainty, Vol 5(4), pp. 297--323. Springer.

43. From anomalies to forecasts: Toward a descriptive model of decisions under risk, under ambiguity, and from experience [link]

    Erev, I., Ert, E., Plonsky, O., Cohen, D. and Cohen, O., 2017. Psychological review, Vol 124(4), pp. 369. American Psychological Association.

44. Cicero: Multi-Turn, Contextual Argumentation for Accurate Crowdsourcing [PDF]
    Chen, Q., Bragg, J., Chilton, L.B. and Weld, D.S., 2018. arXiv preprint arXiv:1810.10733.

45. The rationality of informal argumentation: A Bayesian approach to reasoning fallacies [link]
    Hahn, U. and Oaksford, M., 2007. Psychological review, Vol 114(3), pp. 704. American Psychological Association.

46. Rationality in medical decision making: a review of the literature on doctors' decision-making biases [link]
    Bornstein, B.H. and Emler, A.C., 2001. Journal of evaluation in clinical practice, Vol 7(2), pp. 97--107. Wiley Online Library.

47. Expert political judgment: How good is it? How can we know? [HTML]
    Tetlock, P.E., 2017. Princeton University Press.

48. Two approaches to the study of experts' characteristics [PDF]
    Chi, M.T.H., 2006. The Cambridge Handbook of Expertise and Expert Performance, pp. 21--30.

49. Debiasing [link]
    Larrick, R.P., 2004. Blackwell Handbook of Judgment and Decision Making, pp. 316--338. Wiley Online Library.

50. An evaluation of argument mapping as a method of enhancing critical thinking performance in e-learning environments [link]
    Dwyer, C.P., Hogan, M.J. and Stewart, I., 2012. Metacognition and Learning, Vol 7(3), pp. 219--244. Springer.

51. Forecasting tournaments: Tools for increasing transparency and improving the quality of debate [link]
    Tetlock, P.E., Mellers, B.A., Rohrbaugh, N. and Chen, E., 2014. Current Directions in Psychological Science, Vol 23(4), pp. 290--295. Sage Publications Sage CA: Los Angeles, CA.

52. How to make cognitive illusions disappear: Beyond "heuristics and biases" [link]
Gigerenzer, G., 1991. European review of social psychology, Vol 2(1), pp. 83--115. Taylor & Francis.

53. Liberals and conservatives rely on different sets of moral foundations [PDF]
Graham, J., Haidt, J. and Nosek, B.A., 2009. Journal of personality and social psychology, Vol 96(5), pp. 1029. American Psychological Association.

54. Negative emotions can attenuate the influence of beliefs on logical reasoning [link]
Goel, V. and Vartanian, O., 2011. Cognition and Emotion, Vol 25(1), pp. 121--131. Taylor & Francis.

55. Epistemic democracy: Generalizing the Condorcet jury theorem [link]
List, C. and Goodin, R.E., 2001. Journal of political philosophy, Vol 9(3), pp. 277--306. Wiley Online Library.

56. Aggregating sets of judgments: An impossibility result [PDF]
List, C. and Pettit, P., 2002. Economics \& Philosophy, Vol 18(1), pp. 89--110. Cambridge University Press.

57. The Delphi technique as a forecasting tool: issues and analysis [link]
Rowe, G. and Wright, G., 1999. International Journal of Forecasting, Vol 15(4), pp. 353 - 375. DOI: https://doi.org/10.1016/S0169-2070(99)00018-7

58. OpenAI Charter [link]
OpenAI,, 2018.

Updates and Corrections

If you see mistakes or want to suggest changes, please create an issue on GitHub.

Reuse

Citation

For attribution in academic contexts, please cite this work as

```
Irving & Askell, "AI Safety Needs Social Scientists", Distill, 2019.
```

BibTeX citation

```
@article{irving2019ai,
  author = {Irving, Geoffrey and Askell, Amanda},
  title = {AI Safety Needs Social Scientists},
  journal = {Distill},
  year = {2019},
  note = {https://distill.pub/2019/safety-needs-social-scientists},
  doi = {10.23915/distill.00014}
}
```

 ISSN 2476-0757