# Alignment for Advanced Machine Learning Systems

**Jessica Taylor** and **Eliezer Yudkowsky** and **Patrick LaVictoire** and **Andrew Critch**
Machine Intelligence Research Institute
{jessica,eliezer,patrick,critch}@intelligence.org

**Abstract**

We survey eight research areas organized around one question: As learning systems become increasingly intelligent and autonomous, what design principles can best ensure that their behavior is aligned with the interests of the operators? We focus on two major technical obstacles to AI alignment: the challenge of specifying the right kind of objective functions, and the challenge of designing AI systems that avoid unintended consequences and undesirable behavior even in cases where the objective function does not line up perfectly with the intentions of the designers.

Open problems surveyed in this research proposal include: How can we train reinforcement learners to take actions that are more amenable to meaningful assessment by intelligent overseers? What kinds of objective functions incentivize a system to "not have an overly large impact" or "not have many side effects"? We discuss these questions, related work, and potential directions for future research, with the goal of highlighting relevant research topics in machine learning that appear tractable today.

## Contents

## 1 Introduction

Recent years' progress in artificial intelligence has prompted renewed interest in a question posed by Russell and Norvig (2010): "What if we succeed?" If and when AI researchers succeed at the goal of designing machines with cross-domain learning and decision-making capabilities that rival those of humans, the consequences for science, technology, and human life are likely to be large.

For example, suppose a team of researchers wishes to use an advanced ML system to generate plans for finding a cure for Parkinson's disease. They might approve if it generated a plan for renting computing resources to perform a broad and efficient search through the space of remedies. They might disapprove if it generates a plan to proliferate robotic laboratories which would perform rapid and efficient experiments, but have a large negative effect on the biosphere. The question is, how can we design systems (and select objective functions) such that our ML systems reliably act more like the former case and less like the latter case?

Intuitively, it seems that if we could codify what we mean by "find a way to cure Parkinson's disease *without doing anything drastic*," many of the dangers Bostrom (2014) describes in his book *Superintelligence* could be ameliorated. However, naïve attempts to formally specify satisfactory objectives for this sort of goal usually yield functions that, upon inspection, are revealed to incentivize unintended behavior. (For examples, refer to Soares et al. [2015] and Armstrong [2015].)

What are the key technical obstacles here? Russell (2014) highlights two: a system's objective function "may not be perfectly aligned with the values of the human race, which are (at best) very difficult to pin down;" and "any sufficiently capable intelligent system will prefer to ensure its own continued existence and to acquire physical and computational resources—not for their own sake, but to succeed in its assigned task." In other words, there are at least two obvious types of research that would improve the ability of researchers to design aligned AI systems in the future: We can do research that makes it easier to specify our intended goals as objective functions; and we can do research aimed at designing AI systems that avoid large side effects and negative incentives, even in cases where the objective function is imperfectly aligned. Soares and Fallenstein (2014) refer to the former approach as *value specification*, and the latter as *error tolerance*.

In this document, we explore eight research areas based around these two approaches to aligning advanced ML systems, many of which are already seeing interest from the larger ML community. Some focus on value specification, some on error tolerance, and some on a mix of both. Since reducing the risk of catastrophe from fallible human programmers is itself a shared human value, the line between these two research goals can be blurry.

For solutions to the problems discussed below to be useful in the future, they must be applicable even to ML systems that are much more capable than the systems that exist today. Solutions that critically depend on the system's ignorance of a certain discoverable fact, or on its inability to come up with a particular strategy, should be considered unsatisfactory in the long term. As discussed by Christiano (2015c), if the techniques used to align ML systems with their designers' intentions cannot scale with intelligence, then large gaps will emerge between what we can safely achieve with ML systems and what we can *efficiently* achieve with ML systems.

We will focus on safety guarantees that may seem extreme in typical settings where ML is employed today, such as guarantees of the form, "After a certain period, the system makes zero significant mistakes." These sorts of guarantees are indispensable in safety-critical systems, where a small mistake can have catastrophic real-world consequences. (Guarantees of this form have precedents, e.g., in the KWIK learning framework of Li, Littman, and Walsh [2008].) We will have these sorts of strong guarantees in mind when we consider toy problems and simple examples.

The eight research topics we consider are:

1. **Inductive ambiguity identification:** How can we train ML systems to detect and notify us of cases where the classification of test data is highly under-determined from the training data?

2. **Robust human imitation:** How can we design and train ML systems to effectively imitate humans who are engaged in complex and difficult tasks?

3. **Informed oversight:** How can we train a reinforcement learning system to take actions that aid an intelligent overseer, such as a human, in accurately assessing the system's performance?

4. **Generalizable environmental goals:** How can we create systems that robustly pursue goals defined in terms of the state of the environment, rather than defined directly in terms of their sensory data?

5. **Conservative concepts:** How can a classifier be trained to develop useful concepts that exclude highly atypical examples and edge cases?

6. **Impact measures:** What sorts of regularizers incentivize a system to pursue its goals with minimal side effects?

7. **Mild optimization:** How can we design systems that pursue their goals "without trying too hard", i.e., stopping when the goal has been pretty well achieved, as opposed to expending further resources searching for ways to achieve the absolute optimum expected score?

8. **Averting instrumental incentives:** How can we design and train systems such that they robustly lack default incentives to manipulate and deceive the operators, compete for scarce resources, etc.?

In Section 2, we briefly introduce each topic in turn, alongside samples of relevant work in the area. We then discuss directions for further research that we expect to yield tools which would aid in the design of ML systems that would be robust and reliable, given large amounts of capability, computing resources, and autonomy.

### 1.1 Motivations

In recent years, progress in the field of machine learning has advanced by leaps and bounds. Xu et al. (2015) used an attention-based model to evaluate and describe images (via captions) with remarkably high accuracy. Mnih et al. (2016) used deep neural networks and reinforcement learning to achieve good performance across a wide variety of Atari games. Silver et al. (2016) used deep networks trained via both supervised and reinforcement learning and paired with Monte-Carlo simulation techniques, to beat the human world champion at Go. Lake, Salakhutdinov, and Tenenbaum (2015) use hierarchical Bayesian models to learn visual concepts using only a single example.

In the long run, computer systems making use of machine learning and other AI techniques will become more and more capable, and humans will likely trust those systems to make larger decisions and greater autonomy. As the capabilities of these systems increase, it becomes ever-more important that they act in accordance with the intentions of their operators, and without posing risks to society at large.

As AI systems gain in capability, it will become more difficult to design training procedures and test regimes that reliably align those systems with the intended goals. As an example, consider the task of training a reinforcement learner to play video games by rewarding it according to its score (as per Mnih et al. [2013]). If the learner were to find glitches in the game that allow it to get very high scores, it would switch to a strategy of exploiting those glitches and ignore the features of the game that the programmers are interested in. Somewhat counter-intuitively, improving systems' capabilities can make them *less* likely to "win the game" in the sense we care about, because smarter systems can better find loopholes in training procedures and test regimes. (For a simple example of this sort of behavior with a fairly weak reinforcement learner, refer to Murphy [2013].)

Intelligent systems' capacity to solve problems in surprising ways is a feature, not a bug. One of the key attractions of learning systems is that they can find clever ways to meet objectives that their programmers wouldn't have thought of.

However, this property is a double-edged sword: As the system gets better at finding counter-intuitive solutions, it also gets better at finding exploits that allow it to *formally* achieve operators' explicit goals, without satisfying their intended goals.

For intelligent systems pursuing realistic goals in the world, loopholes can be subtler, more abundant, and much more consequential. Consider the challenge of designing robust objective functions for learning systems that are capable of representing facts about their programmers' beliefs and desires. If the programmers learn that the system's objective function is misspecified, then they will want to repair this defect. If the *learner* is aware of this fact, however, then it has a natural incentive to conceal any defects in its objective function, for the system's current objectives are unlikely to be achieved if the system is made to pursue different objectives. (This scenario is discussed in detail by Bostrom [2014] and Yudkowsky [2008]. Benson-Tilsen and Soares [2016] provide a simple formal illustration.)

This motivates the study of tools and methods for specifying objective functions that avert those default incentives, and for developing ML systems that do not "optimize too hard" in pursuit of those objectives.

### 1.2 Relationship to Other Agendas

This list of eight is not exhaustive. Other important research problems bearing on AI's long-term impact have been proposed by Soares and Fallenstein (2014) and Amodei et al. (2016), among others.

Soares and Fallenstein's "Agent Foundations for Aligning Machine Intelligence with Human Interests" (2014), drafted at the Machine Intelligence Research Institute, discusses several problems in value specification (e.g., ambiguity identification) and error tolerance (e.g., corrigibility, a subproblem of averting instrumental incentives). However, that agenda puts significant focus on a separate research program, *highly reliable agent design*. The goal of that line of research is to develop a better general understanding of how to design intelligent reasoning systems that reliably pursue a given set of objectives.

Amodei et al.'s "Concrete Problems in AI Safety" (2016) is, appropriately, more concrete than Soares and Fallenstein or the present agenda. Amodei et al. write that their focus is on "the empirical study of practical safety problems in modern machine learning systems" that are likely to be useful "across a broad variety of potential risks, both short- and long-term." There is a fair amount of overlap between our agenda and Amodei et al.'s; some of the topics in our agenda were inspired by conversations with Paul Christiano, a co-author on the concrete problems agenda. Our approach differs from Amodei et al.'s mainly in focusing on broader and less well-explored topics. We spend less time highlighting areas where we can build on existing research programs, and more time surveying entirely new research directions.

We consider both Soares and Fallenstein's research proposal and Amodei et al.'s to be valuable, as we expect the AI alignment problem to demand theoretical and applied research from a mix of ML scientists and specialists in a number of other disciplines.

For a more general overview of research questions in AI safety, including both strictly near-term and strictly long-term issues in computer science and other disciplines, see Russell, Dewey, and Tegmark (2015).

## 2 Eight Research Topics

In the discussion to follow, we use the term "AI system" when considering computer systems making use of artificial intelligence algorithms in general, usually when considering systems with capabilities that go significantly beyond the current state of the art. We use the term "ML system" when considering computer systems making use of algorithms qualitatively similar to modern machine learning techniques, especially when considering problems that modern ML techniques are already used to solve.

If the system is capable of making predictions (or answering questions) about a rich and complex domain, we will say that the system "has beliefs" about that domain. If the system is optimizing some objective function, we will say that the system "has goals." A system pursuing some set of goals by executing or outputting a series of actions will sometimes be called an "agent."

## 2.1 Inductive Ambiguity Identification

Human values are context-dependent and complex. To have any hope of specifying our values, we will need to build systems that can *learn* what we want inductively (via, e.g., reinforcement learning). To achieve high confidence in value learning systems, however, Soares (2016) argues that we will need to be able to anticipate cases where the system's past experiences of preferred and unpreferred outcomes provide insufficient evidence for inferring whether future outcomes are desirable. More generally, AI systems will need to "keep humans in the loop" and recognize when they are (and aren't) too inexperienced to make a critical decision safely.

Consider a classic parable recounted by Dreyfus and Dreyfus (1992): The US army once built a neural network intended to distinguish between Soviet tanks and American tanks. The system performed remarkably well with relatively little training data—so well, in fact, that researchers grew suspicious. Upon inspection, they found that all of the images of Soviet tanks were taken on a sunny day, while the images of US tanks were taken on a cloudy day. The network was discriminating between images based on their brightness, rather than based on the variety of tank depicted.[1]

It is to be expected that a classifier, given training data, will identify very simple boundaries (such as "brightness") that separate the data. However, what we want is a classifier that can, given a data set analogous to the tank training set, recognize that it does not contain any examples of Soviet tanks on cloudy days, and ask the user for clarification. Doing so would likely require larger training sets and different training techniques. The problem of inductive ambiguity identification is to develop robust techniques for automatically identifying this sort of ambiguity and querying the user only when necessary.

**Related work.** Amodei et al. (2016) discuss a very similar problem, under the name of "robustness to distributional change." They focus on the design of ML systems that behave well when the test distribution is different from the training distribution, either by making realistic statistical assumptions that would allow correct generalization, or by detecting the novelty and adopting some sort of conservative behavior (i.e., querying a human). We take the name from Soares and Fallenstein (2014), who call the problem "inductive ambiguity identification." Our framing of the problem differs only slightly from that of Amodei et al. (for instance, they consider "scalable oversight" to be a separate problem, while we place the problem of identifying situations where the training data is insufficient to specify the correct reward function under the umbrella of inductive ambiguity identification), but the underlying technical challenge is the same.

Bayesian approaches to training classifiers (including Bayesian logistic regression [Genkin, Lewis, and Madigan 2007] and Bayesian neural networks [Blundell et al. 2015; Korattikara et al. 2015]) maintain uncertainty over the parameters of the classifier. If such a system has the right variables (such as a variable $L$ tracking the degree to which light levels are relevant to the classification of the tank), such a system could automatically become especially uncertain about instances whose classification depends on unknown variables (such as $L$). The trick is having the right variables (and efficiently maintaining the probability distribution), which is

1. Tom Dietterich relates a similar story (personal conversation, 2016), where in his laboratory, years ago, microscope slides containing different types of bugs were made on different days, and a classifier learned to classify the different types of bugs with remarkably high accuracy—because the sizes of the bubbles in the slides changed depending on the day.

quite difficult in practice. There has been much work studying the problem of feature selection (Liu and Motoda 2007; Guo and Schuurmans 2012), but more work is needed to understand under what conditions Bayesian classifiers will correctly identify important inductive ambiguities.

Non-Bayesian approaches, on the other hand, do not by default identify ambiguities. For example, neural networks are notoriously overconfident in their classifications (Goodfellow, Shlens, and Szegedy 2014; Nguyen, Yosinski, and Clune 2015), and so they do not identify when they should be more uncertain, as illustrated by the parable of the tank classifier. Gal and Ghahramani (2016) have recently made progress on this problem by showing that dropout for neural networks can be interpreted as an approximation to certain types of Gaussian processes.

The field of *active learning* (Settles 2010) also bears on inductive ambiguity identification. Roughly speaking, an active learner will maintain a set of "plausible hypotheses" by, e.g., starting with a certain set of hypotheses and retaining the ones that assigned sufficiently high likelihood to the training data. As long as multiple hypotheses are plausible, some ambiguity remains. To resolve this ambiguity, an active learner will ask the human to label additional images that will rule out some of its plausible hypotheses. For example, in the tank-detection setting, a hypothesis is a mapping from images (of tanks) to probabilities (representing, say, the probability that the tank is a US tank). In this setting, an active learner may synthesize an image of a US tank on a sunny day (or, more realistically, pick one out from a large dataset of unlabeled examples). When the user labels this image as a US tank, the hypothesis that an image contains a US tank if and only if the light level is below a certain threshold is ruled out.

Seung, Opper, and Sompolinsky (1992) and Beygelzimer, Dasgupta, and Langford (2009) have both studied what statistical guarantees can be achieved in this setting. Hanneke (2007) introduced the disagreement coefficient to measure the overall probability of disagreement among a local ball in the concept space under the "probability of disagreement" pseudo-metric, which resembles a notion of "local ambiguity"; the disagreement coefficient has been used to clarify and improve upper bounds on label complexity for active learning algorithms (Hanneke 2014). Beygelzimer et al. (2016) introduced an active learning setting where the learner can request counterexamples to hypotheses, and they showed that this search oracle in some cases can speed up learning exponentially; these results are promising, but to scale to more complex systems, more transparent hypothesis spaces may be necessary for humans to interact efficiently with the learner.

Much work remains to be done. Modern active learning settings usually either assume a very simple hypothesis class, or assume that test examples are independent and identically distributed and are drawn from some distribution that the learner has access to at training time.[2] Both of these assumptions are far too strong for use in the general case, where the set of possible hypotheses is rich and the environment is practically guaranteed to have regularities and dependencies that were not represented in the training data.

For example, consider the case where the data that the ML system encounters during operation depends on the behavior of the system itself—perhaps the Soviets start disguising their tanks (imperfectly) to look like US tanks after learning that the ML system has been deployed. In this case, the assumption that the training data would be similar to the test data is violated, and the guarantees disappear. This phenomenon is already seen in certain adversarial settings, such as when spammers change their spam messages in response to how spam recognizers work. Guaranteeing good behavior when the test data differs from the training data is the subject of research in the adversarial machine learning subfield (see, e.g., Huang et al. [2011]). It will take a fair bit of effort to apply those techniques to the active learning setting.

Conformal prediction (Vovk, Gammerman, and Shafer 2005) is an alternative non-Bayesian approach that attempts to produce well-calibrated predictions. In

2. Some forms of online active learning (refer to, e.g. Dekel, Gentile, and Sridharan [2012]) relax the i.i.d. assumption, but the authors do not see how to apply them to the problem of inductive ambiguity identification.

an online classification setting, a conformal predictor will give a *set* of plausible classifications for each instance, and under certain exchangeability assumptions, this set will contain the true classification about (say) 95% of the time throughout the online learning process. This will detect ambiguities in the sense that the conformal predictor must usually output a set containing multiple different classifications for ambiguous instances, on pain of failing to be well-calibrated. However, the exchangability assumption used in conformal prediction is only slightly weaker than an i.i.d. assumption, and the well-calibrated confidence regions (such as 95% true classification) are insufficient for our purposes (where even a single error could be highly undesirable).

KWIK ("Knows What It Knows") learning (Li, Littman, and Walsh 2008) is a variant of active learning that relaxes the i.i.d. assumption, queries the humans only finitely many times, and (under certain conditions) makes *zero* critical errors. Roughly speaking, the KWIK learning framework is one where a learner maintains a set of "plausible hypotheses" and makes classifications only when all remaining plausible hypotheses agree on how to do so. If there is significant disagreement among the plausible hypotheses, a KWIK learner will output a special value $\perp$ indicating that the classification is ambiguous (at which point a human can provide the correct label for that input). The KWIK framework is concerned with algorithms that are guaranteed to output $\perp$ only a limited number of times (usually polynomial in the dimension of the hypothesis space). This guarantees that the system eventually has good behavior, assuming that at least one good hypothesis remains plausible. In the tank classification problem, if the system had a hypothesis for "the user cares about tank type" and another for "the user cares about brightness," then, upon finding a bright picture of a US tank, the system would output $\perp$ and require a human to provide a label for the ambiguous image.

Currently, efficient KWIK learning algorithms are only known for simple hypothesis classes (such as small finite sets of hypotheses, or low-dimensional sets of linear hypotheses). Additionally, KWIK learning makes a strong realizability assumption: useful statistical guarantees can only be obtained when one of the hypotheses in the set is "correct" in that its probability that the image is classified as a tank is always well-calibrated—otherwise, the right hypothesis might not exist in the "plausible set" (Li, Littman, and Walsh 2008; Khani and Rinard 2016). Thus, significant work needs to be done before these frameworks can be used for the inductive ambiguity identification algorithms of highly capable AI systems operating in the real world.

**Directions for future research.** Further study of Bayesian approaches to classification, including the design of realistic priors, better methods of inferring latent variables, and extensions of Bayesian classification approaches to represent more complex models, could improve our understanding of inductive ambiguity identification.

Another obvious direction for future research is to attempt to extend active learning frameworks, like KWIK, that relax the strong i.i.d. assumption. Research in that direction could include modifications to KWIK that allow more complex hypothesis classes, such as neural networks. This will very likely require making different statistical assumptions than in standard KWIK. What statistical guarantees can be provided in variants of the KWIK framework with weakened assumptions about the complexity of the hypothesis class is an open question.

One could also study different methods of relaxing the realizability assumptions in KWIK learning. An ideal learning procedure will notice when the real world contains patterns that none of its hypotheses can model well and flag its potentially flawed predictions (perhaps by outputting $\perp$) accordingly. The "agnostic KWIK learning framework" of Szita and Szepesvári (2011) handles some forms of nonrealizability, but has severe limitations: even if the hypothesis class is linear, the number of labels provided by the user may be exponential in the number of dimensions of the linear hypothesis class.

Alternatively, note that the standard active learning framework and the KWIK framework both represent inductive ambiguity as disagreement among specific hypotheses that have performed well in the past. This is not the only way to

represent inductive ambiguity; it is possible that some different algorithm will find "natural" ambiguities in the data without representing these ambiguities as disagreements between hypotheses. For example, we could consider systems that use a joint distribution over the answers to all possible queries. Where active learners are uncertain about both which hypothesis is correct *and* what the right answers are given the right hypothesis, a system with a joint would be uncertain only about how to answer queries. In this setting, it may be possible to achieve useful statistical guarantees as long as the distribution contains a grain of truth (i.e., is a mixture between a good distribution and some other distributions). Then, of course, good approximation schema would be necessary, as reasoning according to a full joint would be intractable. Refer to Christiano (2016a) for further discussion of this setup.

### 2.2 Robust Human Imitation

Formally specifying a fully aligned general-purpose objective function by hand appears to be an impossibly difficult task, for reasons that also raise difficulties for specifying a correct value learning process. It is hard to see even in principle how we might attain confidence that the goals an ML system is learning are in fact our true goals, and not a superficially similar set of goals that diverge from our own in some yet-undiscovered cases.

Ambiguity identification can help here, by limiting the agent's autonomy. Inductive ambiguity identifiers suspend their activities to consult with a human operator in cases where training data significantly under-determines the correct course of action. But what if we take this idea to its logical conclusion, and use "consult a human operator for advice" itself as our general-purpose objective function?

The target "do what a trusted human would have done, given some time to think about it" is a plausible candidate for a goal that one might safely and usefully optimize. At least, if optimized correctly, this objective function leads to an outcome no *worse* than what would have occurred if the trusted human had access to the AI system's capabilities (Christiano 2015b).

There are a number of difficulties that arise when attempting to formalize this sort of objective. For example, the formalization itself might need to be designed to avert harmful instrumental strategies such as "performing brain surgery on the trusted human's brain to better figure out what they actually would have done". The high-level question here is: Can we define a measurable objective function for human imitation such that the better a system correctly imitates a human, the better its score according to this objective function?

**Related work.** A large portion of supervised learning research can be interpreted as research that attempts to train machines to imitate the way that humans label certain types of data. Deep neural networks achieve impressive performance on many tasks that require emulating human concepts, such as image recognition (Krizhevsky, Sutskever, and Hinton 2012; He et al. 2015) and image captioning (Karpathy and Fei-Fei 2015). Generative models (as studied by, e.g., Gregor et al. [2015] and Lake, Salakhutdinov, and Tenenbaum [2015]) and imitation learning (e.g., Judah et al. [2014], Ross, Gordon, and Bagnell [2010], and Asfour et al. [2008]) are state-of-the-art when it comes to imitating the behavior of humans in applications where the output space is very large and/or the training data is very limited.

In the inverse reinforcement learning paradigm (Ng and Russell 2000) applied to apprenticeship learning (Abbeel and Ng 2004), the learning system imitates the behavior of a human demonstrator in some task by learning the reward function the human is (approximately) optimizing. Ziebart et al. (2008) used the maximum entropy criterion to convert this into a well-posed optimization problem. Inverse reinforcement learning methods have been successfully applied to autonomous helicopter control, achieving human-level performance (Abbeel, Coates, and Ng 2010), and have recently been extended to the learning of non-linear cost features in the environment, producing good results in robotic control tasks with complicated objectives (Finn, Levine, and Abbeel 2016). IRL methods may not scale safely, however, due to their reliance on the faulty assumption that human demonstrators

are optimizing for a reward function, where in reality humans are often irrational, ill-informed, incompetent, and immoral; recent work by Evans, Stuhlmüller, and Goodman (2015b, 2015a) has begun to address these issues.

These techniques have not yet (to our knowledge) been applied to the high-level question of which human imitation tasks can or can't be performed with some sort of guarantee, and what statistical guarantees are possible, but the topic seems ripe for study.

It is also not yet clear whether imitation of humans can feasibly scale up to complex and difficult tasks (e.g., a human engineer engineering a new type of jet engine, or a topologist answering math questions). For complex tasks, it seems plausible that the system will need to learn a detailed psychological model of a human if it is to imitate one, and that this might be significantly more difficult than training a system to do engineering directly. More research is needed to clarify whether imitation learning can scale efficiently to complex tasks.

**Directions for further research.** To formalize the question of robust human imitation, imagine a system $A$ that answers a series of questions. On each round, it receives a natural language question $x$, and should output a natural language answer $y$ that imitates the sort of answer a particular human would generate. Assume the system has access to a large corpus of training data $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$ of previous questions answered by that human. How can we train $A$ such that we get some sort of statistical guarantee that it eventually robustly generates good answers?

One possible solution lies in the generative adversarial models of Goodfellow et al. (2014), in which a second system $B$ takes answers as input and attempts to tell whether they were generated by a human or by $A$. $A$ can then be trained to generate an answer $y$ that is likely to fool $B$ into thinking that the answer was human-generated. This approach could fail if $B$ is insufficiently capable; for example, if $B$ can understand grammar but not content, then $A$ will only be trained to produce grammatically valid answers (rather than correct answers). Further research is required to understand the limits of this approach.

Variational autoencoders, as described by Kingma and Welling (2013), are a particularly promising approach to training systems that are able to form generative models of their training data, and it might be possible to use variants on those methods to train systems to generate good answers to certain classes of questions (given sufficient training on question/answer pairs). However, it is not yet clear whether variational autoencoder techniques can be used to train systems to imitate humans performing complex tasks. In particular, unlike generative adversarial models (which can, in principle, use arbitrary algorithms to imitate the human), variational autoencoders can only efficiently imitate a human using "reversible" algorithms (Stuhlmüller, Taylor, and Goodman 2013), which is a fairly strict limitation. What exactly is the set of tasks that can be performed using reversible generative models? Can we transform non-reversible tasks into reversible ones by adding appropriate information? Research into these questions and others could help us understand whether (and how) modern autoencoder techniques could be used to train systems to imitate humans when performing complex tasks.

## 2.3 Informed Oversight

One of the reasons why robust human imitation would be valuable is the intuition that the objective function of "do what a trusted human would have approved of, given time to consider" might be relatively easy to formalize in a way that could be optimized without many adverse side effects. This point is argued by Christiano (2015a, 2016b), who refers to such agents as "approval-directed agents." For example, we might train a reinforcement learning system to take actions that a human would rate highly by using a framework where the system has to learn the "human judgment" reward function, and where training data is produced by actually having a human evaluate the learner's actions. This approach is taken by, e.g., the TAMER framework of Knox and Stone (2009), in which a learner myopically pursues short-term human approval.

Whether this sort of approval-directed approach would work well for highly capable ML systems pursuing complex tasks depends on a number of key assumptions, one of which is that a "smarter" agent (such as a human), given time to deliberate, can adequately oversee the operation of a "simpler" agent (such as a reinforcement learner).

At first it might seem like this is easy enough. However, a smarter system probably cannot reliably assess the behavior of a simpler system in full generality. Consider an ML system trained to write original novels, using a corpus of human novels as training data. It might be quite a bit less intelligent than a human (according to many different intelligence metrics), but the human may still have a fair bit of trouble accurately evaluating the system's performance. For instance, it might be easy for the system to simply plagiarize a novel, and it might be hard for a human to check whether the novel was in fact plagiarized. (Perhaps the system used a simple rewriting scheme to make the plagiarism difficult to detect by a text search, but still easy to detect for a human comparing the novels side-by-side.)

How do we make it easy for the human to assess the performance of an advanced ML system pursuing some particular task? As noted by Christiano (2016c), it is not sufficient for the human to be more clever than the system, as some bad actions (such as plagiarism) are easier to execute than they are to detect.

For a somewhat less plausible example in which the difficulty is made particularly clear, imagine that a system $A$ is tasked with generating pictures of cats, and a system $B$ is tasked with ensuring that system $A$ is not sending messages to a third party (system $C$) in the lowest-order bits of the image (which, if system $A$ were generating pictures as intended, would be uniformly random). If both system $A$ and system $B$ know system $C$'s public key, but neither know $C$'s private key, then $B$ has no effective way of distinguishing cat pictures that contain steganographic messages from ones that don't, even if $B$ is more intelligent than $A$.

More important than these specific examples is the general pattern: there are many possible hard-to-detect ways a system's behavior could differ from the intended behavior, and at least some of these differences are undesirable. We would like a general strategy for avoiding problems of this form. How can we train systems to not only take good actions, but take actions that can be accurately assessed by overseers?

**Related work.** As mentioned, the TAMER framework of Knox and Stone (2009) provides an early framework for studying approval-directed agents in a fairly myopic setting. Christiano (2016c) has also discussed this problem in detail. Daniel et al. (2014) extend the TAMER framework with an active learning component, improving over hand-coded reward functions in robot learning tasks. A separate approach to human supervision of ML systems is the cooperative inverse reinforcement learning framework of Hadfield-Menell et al. (2016), which views the human-agent interaction as a cooperative game where both players attempt to find a joint policy that maximizes the human's secret value function. Everitt and Hutter (2016) describe a general value learning agent that avoids some potential problems with reinforcement learning and might reproduce approval-directed behavior given a good understanding of how to learn reward functions. Soares et al. (2015) have considered the question of how to design systems that have no incentive to manipulate or deceive in general.

The informed oversight problem is related to the scalable oversight problem discussed by Amodei et al. (2016), which is concerned with methods for efficiently scaling up the ability of human overseers to supervise ML systems in scenarios where human feedback is expensive. The informed oversight problem is slightly different, in that it focuses on the challenge of supervising ML systems in scenarios where they are complex and potentially deceptive (but where feedback is not necessarily expensive).

We now review some recent work on making ML systems more transparent, which could aid an informed overseer by allowing them to evaluate a system's internal reasons for decisions rather than evaluating the decisions in isolation.

Neural networks are a well-known example of powerful but opaque components of ML systems. Some preliminary techniques have been developed for understand-

ing and visualizing the representations learned by neural networks (Simonyan, Vedaldi, and Zisserman 2013; Zeiler and Fergus 2014; Mahendran and Vedaldi 2015; Goodfellow, Shlens, and Szegedy 2014). Pulina and Tacchella (2010) define coarse abstractions of neural networks that can be more easily verified to satisfy safety constraints, and can be used to generate witnesses to violations of safety constraints.

Ribeiro, Singh, and Guestrin (2016) introduce a method for explaining classifications that finds a sparse linear approximation to the local decision boundary of a given black-box ML system, allowing the human operator to inspect how the classification depends locally on the most important input features; similarly, the method of Baehrens et al. (2010) reports the gradient in the input of the classification judgment. In a related vein, Datta, Sen, and Zick (2016), Štrumbelj and Kononenko (2014), and Robnik-Šikonja and Kononenko (2008) define metrics for reporting the influence of various inputs and sets of inputs on the output of a black-box ML system. It is unclear whether black-box methods will scale to the evaluation of highly capable ML systems.

On the other extreme, opposite to black-box methods, some ML systems are transparent by construction using, e.g., graphical models or dimensionality reduction (Vellido, Martín-Guerrero, and Lisboa 2012). Bayesian networks (Friedman, Geiger, and Goldszmidt 1997; Pearl 2009) have been applied in many domains, including ones where reliability and interpretability are concerns (Weber et al. 2012). The interpretability of matrix factorization models can be improved by replacing them with a Bayesian network that makes similar judgments, without sacrificing too much accuracy (Carmona and Riedel 2015). Janzing et al. (2013) define a framework for quantifying the causal influence between variables in a causal network, which could be used to selectively report only the most causally relevant factor in some judgment. Sparse models offer another approach to ensure transparency, by simplifying the model enough that humans can inspect the entire learned model (if not the training procedure that produced the model). For example, Letham et al. (2015) generate short lists of rules to accurately perform predictive tasks in medicine and in simple games, while maintaining interpretability.

In the realm of decision-making systems, Baraka, Paiva, and Veloso (2015) and Rosenthal, Selvaraj, and Veloso (2016) have begun research into methods for making online robotics systems more transparent to their operators. Thomaz and Breazeal (2006) improved training times for reinforcement learning systems by having the human trainer view the agent's most likely next actions under consideration, and Li et al. (2015) applied similar techniques to the TAMER framework, with mixed results.

Developing more transparent ML systems is a good first step towards enabling humans to oversee complex AI systems, but work to date has not yet addressed the problems of informed oversight that arise when the system is highly capable and may be able to manipulate human supervisors or circumvent their efforts. Significant work remains to be done before we even understand what sort of formal or statistical guarantees we would want in order to justify confidence in our ability to assess a system's behavior.

**Directions for future research.** Christiano (2016c) has suggested training systems to output both an action $a$ and a "report" $r$ intended to help an overseer evaluate the action. In the plagiarism example, the action $a$ would be a story, and the report $r$ could be a list of literary works that the story makes reference to or copies from. We would like the report to be maximally informative, even when this informativeness prevents the action from receiving a high score (such as when the report points out plagiarism). It is not at all clear what the type of $r$ should be in general; i.e., if the system can commit plagiarism, steganography, and some third offence that we did not consider, it is not clear what space of reports is sufficiently open-ended to allow the overseer to detect the suboptimal behavior in all three cases. A plausible research approach is to more clearly understand both what space of reports could be sufficient, and how a system could be trained to optimize the mixed objective of both choosing an action that is evaluated favorably given the reporting

mechanism, and a report that is evaluated as informative given the action.

### 2.4 Generalizable Environmental Goals

Many ML systems have their objectives specified in terms of their sensory data. For example, reinforcement learners have the objective of maximizing discounted reward over time (or, alternatively, minimizing expected/empirical loss), where "reward" and/or "loss" are part of the system's percepts. While these sensory goals can be useful proxies for environmental goals, environmental goals are distinct: Tricking your sensors into perceiving that a sandwich is in the room is not the same as actually having a sandwich in the room.

Let's say that your goal is to design an AI system that directly pursues some environmental goal, such as "ensure that this human gets lunch today." How can we train the system to pursue a goal like that in a manner that is robust against opportunities to interfere with the proxy methods used to specify the goals, such as "the pixels coming from the camera make an image that looks like food"?

If we were training a system to put some food in a room, we might try providing training data by doing things like: placing various objects on a scale in front of a camera, and feeding the data from the camera and the scale into the system, with labels created by humans (which mark the readings from food as good, and the readings from other objects as bad); or having a human in the room press a special button whenever there is food in the room, where button presses are accompanied by reward.

These training data suggest, but do not precisely specify, the goal of placing food in the room. Suppose that the system has some strategy for fooling the camera, the scale, and the human, by producing an object of the appropriate weight that, from the angle of the camera and the angle of the human, looks a lot like a sandwich. The training data provided is not sufficient to distinguish between this strategy, and the strategy of actually putting food in the room.

One way to address this problem is to design more and more elaborate sensor systems that are harder and harder to deceive. However, this is the sort of strategy that is unlikely to scale well to highly capable AI systems. A more scalable approach is to design the system to learn an "environmental goal" such that it would not rate a strategy of "fool all sensors at once" as high-reward, even if it could find such a policy.

**Related work.** Dewey (2011) and Hibbard (2012) have attempted to extend the AIXI framework of Hutter (2005) so that it learns a utility function over world-states instead of interpreting a certain portion of its percepts as a reward primitive.[3] Roughly speaking, these frameworks require programs to specify (1) the type of the world-state; (2) a prior over utility functions (which map world-states to real numbers); and (3) a "value-learning model" that relates utility functions, state-transitions, and observations. If all these are specified, then it is straightforward to specify the ideal agent that maximizes expected utility (through a combination of exploration to learn the utility function, and exploitation to maximize it). This is a good general framework, but significant research remains if we are to have any luck formally specifying (1), (2), and (3).

Everitt and Hutter (2016) make additional progress by showing that in some cases it is possible to specify an agent that will use its reward percepts as evidence about a utility function, rather than as a direct measure of success. While this alleviates the problem of specifying (3) above (the value-learning model), it leaves open the problem of specifying (1), a representation of the state of the world, and (2), a reasonable prior over possible utility functions (such that the agent converges on the goal that the operators actually intended, as it learns more about the world).

3. When the agent is pursuing some objective specified in terms of elements of its own world-model, we call the objective a "utility function," to differentiate this from the case where reward is part of the system's basic percepts. This practice of referring to preferences over world-states as utility functions dates back to von Neumann and Morgenstern (1944).

The problem of generalizable environmental goals is related to the problem of reward hacking, which has been discussed by Dewey (2011) and Amodei et al. (2016), wherein an AI system takes control of the physical mechanism that dispenses reward and alters it. Indeed, the entire reward hacking problem can be seen as stemming from the failure to specify suitable environmental goals.

**Directions for future research.** Suppose the AI system has learned a world-model with state type $\mathcal{S}$ as in model-based reinforcement learning (Heess et al. 2015). We will assume that $\mathcal{S}$ is very high-dimensional, so there is no guarantee that the correct utility function is a simple function of $\mathcal{S}$.

We would like to define a utility function $U$ on $\mathcal{S}$ that returns a high number for states containing a sandwich, and low numbers for states that do not contain a sandwich. To make this problem tractable, we will assume we can identify some goal state $G \in \mathcal{S}$ in which there is certainly a sandwich in the room. This state could be identified by, for example, having the human place a sandwich in the room (as the AI system observes the human), and seeing which state the system thinks the world is in at this point. The system's goal will be to cause the world to be in a state similar to $G$.

To define what it means for some states to be similar to others, we will find a low-dimensional state representation $\phi : \mathcal{S} \to \mathbb{R}^n$ and then define $U(S) := -\|\phi(S) - \phi(G)\|^2$ to measure the distance between the state and $G$. We will defer the question of how $\phi$ should be defined until after discussing an example.

Consider two different possible world-states. In state $A$, the system has just placed a sandwich in the room. In state $B$, the system has placed a realistic image of a sandwich (printed on paper) in front of the camera, placed a rock (with the same weight as a sandwich) on the scale, and tricked the human into pressing the button. To assign a higher utility to state $A$ than state $B$, we must have $\phi(A)$ close to $\phi(G)$ but $\phi(B)$ far from $\phi(G)$. Thus, the state representation $\phi$ must distinguish $A$ from $B$.

While state $A$ and state $B$ predict the same *immediate* observations, they predict different *future* observations given some future actions. For example, if the AI system took the action of moving the camera, in state $A$ it would become clear that the image was printed on paper, while in state $B$ the sandwich would still appear to be a sandwich. It is therefore plausible that, if the system attempts to select $\phi$ so that the future observations following from a state $S$ can be predicted well as a simple function of $\phi(S)$, then $\phi(A)$ and $\phi(B)$ will be significantly different (since they predict different future observations). At this point, it is plausible that the resulting utility function $U$ assigns a higher value to $A$ than $B$.[4]

However, we can consider a third state $C$ that obtains after the AI system unplugs the camera and the scale from its sensors, and plugs in a "delusion box" (a virtual reality world that it has programmed), as discussed by Ring and Orseau (2011). This delusion box could be programmed so that the system's future observations (given arbitrary future actions) are indistinguishable from those that would follow from state $A$. Thus, if $\phi$ is optimized to select features that aid in predicting future observations well, $\phi(C)$ may be very close (or equal) to $\phi(A)$. This would hinder efforts to learn a utility function that assigns high utility to state $A$ but not state $C$. While it is not clear why an AI system would construct this virtual reality world in this example (where putting a sandwich in the room is probably easier that constructing a detailed virtual reality world), it seems more likely that it would if the underlying task is very difficult. (This is the problem of "wireheading," studied by, e.g., Orseau and Ring [2011].)

4. This proposal is related to the work of Abel et al. (2016), who use a state-collapsing function $\phi$ for RL tasks with high-dimensional $\mathcal{S}$. Their agent explores by taking actions in state $A$ that it hasn't yet taken in previous states $B$ with $\phi(B) = \phi(A)$, where $\phi$ maps states to a small set of clusters. They achieve impressive results, suggesting that state-collapsing functions—perhaps mapping to a richer but still low-dimensional representation space—may capture the important structure of an RL task in a way that allows the agent to compare states to the goal state in a meaningful way.

To avoid this problem, it may be necessary to take into account the past leading up to state $A$ or state $C$, rather than just the future starting from these states. Consider the state $C_{t-1}$ that the world is in right before it is in state $C$. In this state, the system has not quite entered the virtual reality world, so perhaps it is able to exit the virtual reality and observe that there is no sandwich on the table. Therefore, state $C_{t-1}$ makes significantly different predictions from state $A$ given some possible future actions. As a result, it is plausible that state $\phi(C_{t-1})$ and $\phi(A)$ are far from each other. Then, if $\phi(C)$ is close to $\phi(A)$, this would imply that $\phi(C_{t-1})$ is far from $\phi(C)$ (by the triangle inequality). Perhaps we can restrict $\phi$ to avoid such large jumps in feature space, so that $\phi(C)$ must be close to $\phi(C_{t-1})$. "Slow" features (such as those detected by $\phi$ under this restriction) have already proved useful in reinforcement learning (Wiskott and Sejnowski 2002), and may also prove useful here. Plausibly, requiring $\phi$ to be slow could result in finding a feature mapping $\phi$ with $\phi(C)$ far from $\phi(A)$, so that $U$ can assign a higher utility to state $A$ than to state $C$.

This approach seems worth exploring, but more work is required to formalize it and study it (both theoretically and empirically).

### 2.5 Conservative Concepts

Many of the concerns raised by Russell (2014) and Bostrom (2014) center on cases where an AI system optimizes some objective, and, in doing so, finds a strange and undesirable edge case. Writes Russell:

> A system that is optimizing a function of $n$ variables, where the objective depends on a subset of size $k < n$, will often set the remaining unconstrained variables to extreme values; if one of those unconstrained variables is actually something we care about, the solution found may be highly undesirable.

We want to be able to design systems that have "conservative" notions of the goals we give them, so they do not formally satisfy these goals by creating undesirable edge cases. For example, if we task an AI system with creating screwdrivers, by showing it 10,000 examples of screwdrivers and 10,000 examples of non-screwdrivers,[5] we might want it to create a pretty average screwdriver as opposed to, say, an extremely tiny screwdriver—even though tiny screwdrivers may be cheaper and easier to produce.

We don't want the system's "screwdriver" concept to be as simple as possible, because the simplest description of "screwdriver" may contain many edge cases (such as very tiny screwdrivers). We also don't want the system's "screwdriver" concept to be perfectly minimal, as then the system may claim that it is unable to produce any new screwdrivers (because the only things it is willing to classify as screwdrivers are the 10,000 training examples it actually saw, and it cannot perfectly duplicate any of those to the precision of the scan). Rather, we want the system to have a conservative notion of what it means for something to be a screwdriver, such that we can direct it to make screwdrivers and get a sane result.

**Related work.** The naïve approach is to train a classifier to distinguish positive examples from negative examples, and then have it produce an object which it classifies as a positive instance with as high confidence as possible. Goodfellow, Shlens, and Szegedy (2014) have noted that systems trained in this way are vulnerable to exactly the sort of edge cases we are trying to avoid. In training a classifier, it is important that the negative examples given as training data are representative of the negative examples given during testing. But when optimizing the probability the classifier assigns to an instance, the relevant negative examples (edge cases) are often not represented well in the training set. While some work has been done to train systems on these "adversarial" examples, this does not yet resolve the

5. In the simplest case, we can assume that these objects are specified as detailed 3D scans. If we have only incomplete observations of these objects, problems described in Section 2.4 arise.

problem. Resisting adversarial examples requires getting correct labels for many "weird" examples (which humans may find difficult to judge correctly), and even after including many correctly-labeled adversarial examples in the training set, many models (including current neural networks) will still have additional adversarial examples.

Inverse reinforcement learning (Ng and Russell 2000) provides a second method for learning intended concepts, but runs into some of the same difficulties. Naïve approaches to reinforcement learning would allow a learner to distinguish between positive and negative examples of a concept, but would still by default learn a simple separation of the concepts, such that maximizing the learned reward function would likely lead the system towards edge cases.

A third obvious approach is generative adversarial modeling, as studied by Goodfellow et al. (2014). In this framework, one system (the "actor") can attempt to create objects similar to positive examples, while another (the "critic") attempts to distinguish those objects from actual positive examples in the training set. Unfortunately, for complex tasks it may be infeasible in practice to synthesize instances that are statistically indistinguishable from the elements of the training set, because the system's ability to distinguish different elements may far exceed its ability to synthesize elements with high precision. (In the screwdriver case, imagine that the AI system does not have access to any of the exact shades of paint used in the training examples.)

Many of these frameworks would likely be usefully extended by good anomaly detection, which is currently being studied by Siddiqui et al. (2016) among others.

**Directions for future research.** One additional obvious approach to training conservative concepts is to use dimensionality reduction (Hinton and Salakhutdinov 2006) to find the important features of training instances, then use generative models to synthesize new examples that are similar to the training instances only with respect to those specific features. It is not yet clear that this thwarts the problem of edge cases; if the dimensionality reduction were done via autoencoder, for example, the autoencoder itself may beget adversarial examples ("weird" things that it declares match the training data on the relevant features). Good anomaly detection could perhaps ameliorate some of these concerns. One plausible research path is to apply modern techniques for dimensionality reduction and anomaly detection, probe the limitations of the resulting system, and consider modifications that could resolve these problems.

Techniques for solving the inductive ambiguity identification problem (discussed in Section 2.1) could also help with the problem of conservative concepts. In particular, the conservative concept could be defined to be the set of instances that are considered *unambiguously* positive.

At the moment, it is not yet entirely clear what counts as a "reasonable" conservative concept, nor even whether "conservative concepts" (that is, concepts which are neither maximally small nor maximally simple, but which instead match our intuitions about conservatism) are a natural kind. Much of the above research could be done with the goal in mind of developing a better understanding of what counts as a good "conservative concept."

## 2.6 Impact Measures

We would prefer a highly intelligent AI system to avoid creating large unintended-by-us side effects in pursuit of its objectives, and also to notify us of any large impacts that might result from achieving its goal. For example, if we ask it to build a house for a homeless family, it should know implicitly that it should avoid destroying nearby houses for materials—a large side effect. However, we cannot simply design it to avoid having large effects in general, since we would like the system's actions to still have the desirable large follow-on effect of improving the family's socioeconomic situation. For any specific task, we can specify ad-hoc cost functions for side effects like the destruction of nearby houses, but since we cannot always anticipate such costs in advance, we want a quantitative understanding of how to generally limit

an AI systems' side effects (without also limiting its ability to have large positive intended impacts).

The goal of research towards a low-impact measure would be to develop a regularizer on the actions of an AI system that penalizes "unnecessary" large side effects (such as stripping materials from nearby houses) but not "intended" side effects (such as someone getting to live in the house).

**Related work.** Amodei et al. (2016) discuss the problem of impact measures, and describe a number of methods for defining, learning, and penalizing impact in order to incentivize RL agents to steer clear of negative side effects (such as penalizing empowerment, as formalized by Salge, Glackin, and Polani [2014]). However, each of the methods they propose has significant drawbacks (which they describe).

Armstrong and Levinstein (2015) discuss a number of ideas for impact measures that could be used to design objective functions that penalize impact. The general theme is to define a special null policy $\varnothing$ and a variable $V$ that summarizes the state of the world (as best the system can predict it) down into a few key features. (Armstrong suggests having those features be hand-selected, but they could plausibly also be generated from the system's own world-model.) The impact of the policy $\pi$ can then be measured by looking at the divergence between the distribution of $V$ if the system executes $\pi$, compared to the distribution of $V$ if it executes $\varnothing$, with divergence measured as by, e.g., earth mover's distance (Rubner, Tomasi, and Guibas 2000). To predict which state results from each policy, the system must learn a state transition function; this could be done using, e.g., model-based reinforcement learning (Heess et al. 2015).

The main problem with this proposal is that it cannot separate intended follow-on effects from unintended side effects. Suppose a system is given the goal of constructing a house for the operator while having a low impact. Normally, constructing the house would allow the operator to live in the house for some number of years, possibly having effects on the operator, the local economy, and the operator's career. This would be considered an impact under, e.g., the earth mover's distance. Therefore, perhaps the system can get a lower impact score by building the house while preventing the operator from entering it. This limitation will become especially problematic if we plan to use the system to accomplish large-scale goals, such as curing cancer.

**Directions for future research.** It may be possible to use the concept of a causal counterfactual (as formalized by Pearl [2000]) to separate some intended effects from some unintended ones. Roughly, "follow-on effects" could be defined as those that are causally downstream from the achievement of the goal of building the house (such as the effect of allowing the operator to live somewhere). Follow-on effects are likely to be intended and other effects are likely to be unintended, although the correspondence is not perfect. With some additional work, perhaps it will be possible to use the causal structure of the system's world-model to select a policy that has the follow-on effects of the goal achievement but few other effects.

Of course, it would additionally be desirable to query the operator about possible effects, in order to avoid unintended follow-on effects (such as the house eventually collapsing due to its design being structurally unsound) and allow tolerable non-follow-on effects (such as spending money on materials). Studying ways of querying the operator about possible effects this way might be another useful research avenue for the low impact problem.

## 2.7 Mild Optimization

Many of the concerns discussed by Bostrom (2014) in the book *Superintelligence* describe cases where an advanced AI system is maximizing an objective *as hard as possible*. Perhaps the system was instructed to make paperclips, and it uses every resource at its disposal and every trick it can come up with to make literally as many paperclips as is physically possible. Perhaps the system was instructed to

make only 1000 paperclips, and it uses every resource at its disposal and every trick it can come up with to make sure that it *definitely* made 1000 paperclips (and that its sensors didn't have any faults). Perhaps an impact measure was used to penalize side effects, and it uses every resource at its disposal to (as discreetly as possible) prevent bystanders from noticing it as it goes about its daily tasks.

In all of these cases, intuitively, we want some way to have the AI system just "not try so hard." It should expend enough resources to achieve its goals pretty well, with pretty high probability, using plans that are clever enough but not "maximally clever." The problem of mild optimization is: how can we design AI systems and objective functions that, in this intuitive sense, don't optimize more than they have to?

Many modern AI systems are "mild optimizers" simply due to their lack of resources and capabilities. As AI systems improve, it becomes more and more difficult to rely on this method for achieving mild optimization. As noted by Russell (2014), the field of AI is classically concerned with the goal of maximizing the extent to which automated systems achieve some objective. Developing formal models of AI systems that "try as hard as necessary but no harder" is an open problem, and may require significant research.

**Related work.** Regularization (as a general tool) is conceptually relevant to mild optimization. Regularization helps ML systems prevent overfitting, and has been applied to the problem of learning value functions for policies in order to learn less-extreme policies that are more likely to generalize well (Farahmand et al. 2009). It is not yet clear how to regularize algorithms against "optimizing too hard," because it is not yet clear how to measure optimization. There do exist metrics for measuring something like optimization capability (such as the "universal intelligence metric" of Legg and Hutter [2007] and the empowerment metric for information-theoretic entanglement of Klyubin, Polani, and Nehaniv [2005] and Salge, Glackin, and Polani [2014]), but to our knowledge, no one has yet attempted to regularize *against* excessive optimization.

Early stopping, wherein an algorithm is terminated prematurely in attempts to avoid overfitting, is an example of ad-hoc mild optimization. A learned function that is over-optimized just for accuracy on the training data would generalize less well than if it were less optimized. (For a discussion of this phenomenon, refer to Yao, Rosasco, and Caponnetto [2007] and Hinton et al. [2012]).

To make computer games more enjoyable, AI players are often restricted in the amount of optimization pressure (such as search depth) they can apply to their choice of action (Rabin 2010), especially in domains like chess where efficient AI players are vastly superior to human players. We can view this as a response to the fact that the actual goal ("challenge the human player, but not too much") is quite difficult to specify.

Bostrom (2014) has suggested that we design agents to satisfice expected reward, in the sense of Simon (1956), instead of maximizing it. This would work fine if the system found "easy" strategies before finding extreme strategies. However, that may not always be the case: If you direct a clever system to make at least 1,234,567 paper clips, with a satisficing threshold of 99.9% probability of success, the first strategy it considers might be "make as many paper clips as is physically possible," and this may have more than a 99.9% chance of success (a flaw that Bostrom acknowledges).

Taylor (2015) suggests an alternative, which she calls "quantilization." Quantilizers select their action randomly from the top (say) 1% of their possible actions (under some measure), sorted by probability of success. Quantilization can be justified by certain adversarial assumptions: if there is some unknown cost function on actions, and this cost function is the *least convenient* possible cost function that does not assign much expected cost to the average action, then quantilizing is the optimal strategy when maximizing expected reward and minimizing expected cost. The main problem with quantilizers is that it is difficult to define an appropriate measure over actions, one such that a random action in the top 1% of this measure will likely solve the task, but sampling a random action according to that measure is still safe. However, quantilizers point in a promising direction: perhaps it is possible to

make mild optimization part of the AI system's goal, by introducing appropriate adversarial assumptions.

**Directions for future research.** Mild optimization is a wide-open field of study. One possible first step would be to investigate whether there is a way to design a regularizer that penalizes systems for displaying high intelligence (relative to some intelligence metric) in a manner that causes them to achieve the goal quickly and with few wasted resources, as opposed to simply making the system behave in a less intelligent fashion.

Another approach would be to design a series of environments similar to the environment of a classic Atari game, in which the environment contains glitches and bugs that could be exploited via some particularly clever sequence of actions. This would provide a testing environment in which different methods of designing systems that get a high score while refraining from using the glitches and bugs could be tested and evaluated (with an eye towards algorithms that do so in a fashion that is likely to generalize).

Another avenue for future research is to explore and extend the quantilization framework of Taylor (2015) to work in settings where the action measure is difficult to specify.

Research into averting instrumental incentives (discussed below) could help us understand how to design systems that do not attempt to self-modify or outsource computation to the physical world. This would simplify the problem greatly, as it might then be possible to tune a system's capabilities until it is only able to achieve good-enough results, without worrying that the system would simply acquire more resources (and start maximizing in a non-mild manner) given the opportunity to do so.

## 2.8 Averting Instrumental Incentives

Omohundro (2008) has noted that highly capable AI systems should be expected to pursue certain convergent instrumental strategies, such as preservation of the system's current goals and the acquisition of resources. Omohundro's argument is that most objectives imply that an agent pursuing the objective should (1) ensure nobody redirects the agent towards different objectives, as then the current objective would not be achieved; (2) ensure that the agent is not destroyed, as then the current objective would not be achieved; (3) become more resource-efficient; (4) acquire more resources, such as computing resources and energy sources; and (5) improve cognitive capacity.

It is difficult to define practical objective functions that resist these pressures (Benson-Tilsen and Soares 2016). For example, if the system is rewarded for shutting down when the humans want it to shut down, then the system has incentives to take actions that make the humans want to shut it down (Armstrong 2010).

A number of "value learning" proposals, such as those discussed by Hadfield-Menell et al. (2016) and Soares (2016), describe systems that would avert instrumental incentives by dint of the system's uncertainty about which goal it is supposed to optimize. A system that believes that the operators (and only the operators) possess knowledge of the "right" objective function might be very careful in how it deals with the operators, and this caution could counteract potentially harmful default incentives.

This, however, is not the same as *eliminating* those incentives. If a value learning system were ever confidently wrong, the standard instrumental incentives would re-appear immediately. For instance, if the value learning framework were set up slightly incorrectly, and the system gained high confidence that humans terminally value the internal sensation of pleasure, it might acquire strong incentives to acquire a large amount of resources that it could use to put as many humans as possible on opiates.

If we could design objective functions that averted these default incentives, that would be a large step towards answering the concerns raised by Bostrom (2014) and

others, many of which stem from the fact that these subgoals naturally arise from almost any goal.

**Related work.** Soares et al. (2015) and Orseau and Armstrong (2016) have worked on specific designs that can avert specific instrumental incentives, such as the incentive to manipulate a shutdown button or the incentive to avoid being interrupted. However, these approaches have major shortcomings (discussed in those papers), and a satisfactory solution will require more research.

Where those authors pursue methods for averting specific instrumental pressures (namely, pressure to avoid being shut down), it is possible that there may be a general solution to problems of this form, which can be used to simultaneously avert numerous instrumental pressures (including, e.g., the incentive to outsource computation to the environment). Given that a general-purpose method for averting all instrumental pressures (both foreseen and unforeseen) would make it significantly easier to justify confidence that an AI system will behave in a robustly beneficial manner, this topic of research seems well worth pursuing.

**Directions for future research.** Soares et al. (2015), Armstrong (2010), and Orseau and Armstrong (2016) study methods for combining objective functions in such a way that the humans have the ability to switch which function an agent is optimizing, but the agent does not have incentives to cause or prevent this switch. All three approaches leave much to be desired, and further research along those paths seems likely to be fruitful.

In particular, we would like a way of combining objective functions such that the AI system (1) has no incentive to cause or prevent a shift in objective function; (2) is incentivized to preserve its ability to update its objective function in the future; and (3) has reasonable beliefs about the relation between its actions and the mechanism that causes objective function shifts. We do not yet know of a solution that satisfies all of these desiderata. Perhaps a solution to this problem will generalize to also allow the creation of an AI system that also has no incentive to change, for example, the amount of computational resources it has access to.

Another approach is to consider creating systems that "know they are flawed" in some sense. The idea would be that the system would want to shut down as soon as it realizes that humans are attempting to shut it down, on the basis that humans are less flawed than it is. It is difficult to formalize such an idea; naïve attempts result in a system that attempts to model the different ways it could be flawed and optimize according to a mixture over all different ways it could be flawed, which is problematic if the model of various possible flaws is *itself* flawed. While it is not at all clear how to make the desired type of reasoning more concrete, success at formalizing it could result in entirely new approaches to the problem of averting instrumental incentives.

## 3 Conclusion

A better understanding of any of the eight open research areas described above would improve our ability to design robust and reliable AI systems in the future. To review:

1,2,3—A better understanding of robust inductive ambiguity identification, human imitation, and informed oversight would aid in the design of systems that can be safely overseen by human operators (and which query the humans when necessary).

4—Better methods for specifying environmental goals would make it easier to design systems that are pursuing the objectives that we actually care about.

5,6,7—A better understanding of conservative concepts, low-impact measures, and mild optimization would make it easier to design highly advanced systems that fail gracefully and admit of online testing and modification. A conservative, low-impact, mildly-optimizing superintelligent system would be much easier to safely use than a superintelligence that attempts to literally maximize a particular objective function.

8—A general-purpose strategy for averting convergent instrumental subgoals would help us build systems that avert undesirable default incentives such as incentives to deceive their operators and compete for resources.

In working on problems like those discussed above, it is important to keep in mind that they are intended to address whatever long-term concerns with highly intelligent systems we can predict in advance. Solutions that work for modern systems but would predictably fail for highly capable systems are unsatisfactory, as are solutions that work in theory but are prohibitively expensive in practice.

These eight areas of research help support the claim that there are open technical problems—some of which are already receiving a measure of academic attention—whose investigation is likely to be helpful down the road for practitioners attempting to actually build robustly beneficial advanced ML systems.

**Acknowledgments.** Thanks to Paul Christiano for seeding many of the initial ideas for these research directions (and, to a lesser extent, Dario Amodei and Chris Olah). In particular, the problems of informed oversight and robust human imitation were both strongly influenced by Paul. Thanks to Nate Soares and Tsvi Benson-Tilsen for assisting in the presentation of this paper. Thanks to Stuart Armstrong for valuable discussion about these research questions, especially the problem of averting instrumental incentives. Thanks also to Jan Leike, Owain Evans, Stuart Armstrong, and Jacob Steinhardt for valuable conversations.

## References

Abbeel, Pieter, Adam Coates, and Andrew Ng. 2010. "Autonomous helicopter aerobatics through apprenticeship learning." *The International Journal of Robotics Research.*

Abbeel, Pieter, and Andrew Y. Ng. 2004. "Apprenticeship Learning via Inverse Reinforcement Learning." In *21st International Conference on Machine Learning (ICML-'04).* Banff, AB, Canada: ACM. `http://doi.acm.org/10.1145/1015330.1015430`.

Abel, David, Alekh Agarwal, Akshay Krishnamurthy Fernando Diaz, and Robert E. Schapire. 2016. "Exploratory Gradient Boosting for Reinforcement Learning in Complex Domains." In *Abstraction in Reinforcement Learning workshop at ICML-'16.* New York, NY.

Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. "Concrete Problems in AI Safety." arXiv: `1606.06565 [cs.AI]`.

Armstrong, Stuart. 2010. *Utility Indifference,* Technical Report 2010-1. Oxford: Future of Humanity Institute, University of Oxford. `http://www.fhi.ox.ac.uk/utility-indifference.pdf`.

———. 2015. "Motivated Value Selection for Artificial Agents." In *1st International Workshop on AI and Ethics at AAAI-2015.* Austin, TX.

Armstrong, Stuart, and Benjamin Levinstein. 2015. "Reduced Impact Artificial Intelligences." Unpublished draft, `https://dl.dropboxusercontent.com/u/23843264/Permanent/Reduced_impact_S+B.pdf`.

Asfour, Tamim, Pedram Azad, Florian Gyarfas, and Rüdiger Dillmann. 2008. "Imitation learning of dual-arm manipulation tasks in humanoid robots." *International Journal of Humanoid Robotics* 5 (02): 183–202.

Baehrens, David, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. 2010. "How to explain individual classification decisions." *Journal of Machine Learning Research* 11 (Jun): 1803–1831.

Baraka, Kim, Ana Paiva, and Manuela Veloso. 2015. "Expressive Lights for Revealing Mobile Service Robot State." In *Robot'2015, the 2nd Iberian Robotics Conference.* Lisbon, Portugal.

Benson-Tilsen, Tsvi, and Nate Soares. 2016. "Formalizing Convergent Instrumental Goals." In *2nd International Workshop on AI, Ethics and Society at AAAI-2016.* Phoenix, AZ.

Beygelzimer, Alina, Sanjoy Dasgupta, and John Langford. 2009. "Importance Weighted Active Learning." In *Proceedings of the 26th Annual International Conference on Machine Learning,* 49–56. ICML '09. Montreal, Quebec, Canada: ACM. ISBN: 978-1-60558-516-1. doi:10.1145/1553374.1553381. http://doi.acm.org/10.1145/1553374.1553381.

Beygelzimer, Alina, Daniel Hsu, John Langford, and Chicheng Zhang. 2016. "Search Improves Label for Active Learning." *arXiv preprint arXiv:1602.07265.*

Blundell, Charles, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. 2015. "Weight Uncertainty in Neural Networks." arXiv: 1505.05424 [stat.ML].

Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies.* New York: Oxford University Press.

Carmona, Iván Sánchez, and Sebastian Riedel. 2015. "Extracting Interpretable Models from Matrix Factorization Models." In *NIPS.*

Christiano, Paul. 2015a. "Abstract approval-direction." November 28. https://medium.com/ai-control/abstract-approval-direction-dc5a3864c092.

———. 2015b. "Mimicry and Meeting Halfway." September 19. https://medium.com/ai-control/mimicry-maximization-and-meeting-halfway-c149dd23fc17.

———. 2015c. "Scalable AI control." December 5. https://medium.com/ai-control/scalable-ai-control-7db2436feee7.

———. 2016a. "Active Learning for Opaque, Powerful Predictors." January 3. https://medium.com/ai-control/active-learning-for-opaque-powerful-predictors-94724b3adf06.

———. 2016b. "Approval-directed algorithm learning." February 21. https://medium.com/ai-control/approval-directed-algorithm-learning-bf1f8fad42cd.

———. 2016c. "The Informed Oversight Problem." April 1. https://medium.com/ai-control/the-informed-oversight-problem-1b51b4f66b35.

Daniel, Christian, Malte Viering, Jan Metz, Oliver Kroemer, and Jan Peters. 2014. "Active reward learning." In *Proceedings of Robotics Science & Systems.*

Datta, Anupam, Shayak Sen, and Yair Zick. 2016. "Algorithmic Transparency via Quantitative Input Influence." In *Proceedings of 37th IEEE Symposium on Security and Privacy.*

Dekel, Ofer, Claudio Gentile, and Karthik Sridharan. 2012. "Selective Sampling and Active Learning from Single and Multiple Teachers." *Journal of Machine Learning Research* 13 (1): 2655–2697.

Dewey, Daniel. 2011. "Learning What to Value." In Schmidhuber, Thórisson, and Looks 2011, 309–314.

Dreyfus, Hubert L., and Stuart E. Dreyfus. 1992. "What Artificial Experts Can and Cannot Do." *AI & Society* 6 (1): 18–26.

Evans, Owain, Andreas Stuhlmüller, and Noah Goodman. 2015a. "Learning the Preferences of Bounded Agents." *Abstract for NIPS 2015 Workshop on Bounded Optimality.* http://web.mit.edu/owain/www/nips-workshop-2015-website.pdf.

———. 2015b. "Learning the Preferences of Ignorant, Inconsistent Agents." *CoRR* abs/1512.05832. http://arxiv.org/abs/1512.05832.

Everitt, Tom, and Marcus Hutter. 2016. "Avoiding wireheading with value reinforcement learning." arXiv: 1605.03143.

Farahmand, Amir M., Mohammad Ghavamzadeh, Csaba Szepesvári, and Shie Mannor. 2009. "Regularized Policy Iteration." In *Advances in Neural Information Processing Systems 21 (NIPS 2008),* edited by D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, 441–448. Curran Associates, Inc.

Finn, Chelsea, Sergey Levine, and Pieter Abbeel. 2016. "Guided Cost Learning: Deep Inverse Optimal Control via Policy Optimization." *arXiv preprint arXiv:1603.00448.*

Friedman, Nir, Dan Geiger, and Moises Goldszmidt. 1997. "Bayesian network classifiers." *Machine learning* 29 (2-3): 131–163.

Gal, Yarin, and Zoubin Ghahramani. 2016. "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning."

Genkin, Alexander, David D. Lewis, and David Madigan. 2007. "Large-Scale Bayesian Logistic Regression for Text Categorization." *Technometrics* 49 (3): 291–304.

Goodfellow, Ian J., Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. "Generative Adversarial Networks." arXiv: `1406.2661 [stat.ML]`.

Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. 2014. "Explaining and Harnessing Adversarial Examples." arXiv: `1412.6572 [stat.ML]`.

Gregor, Karol, Ivo Danihelka, Alex Graves, and Daan Wierstra. 2015. "DRAW: A recurrent neural network for image generation." arXiv: `1502.04623 [cs.CV]`.

Guo, Yuhong, and Dale Schuurmans. 2012. "Convex structure learning for Bayesian networks: Polynomial feature selection and approximate ordering." *arXiv preprint arXiv:1206.6832.*

Hadfield-Menell, Dylan, Anca Dragan, Pieter Abbeel, and Stuart Russell. 2016. "Cooperative Inverse Reinforcement Learning." arXiv: `1606.03137 [cs.AI]`.

Hanneke, Steve. 2007. "A bound on the label complexity of agnostic active learning." In *Proceedings of the 24th international conference on Machine learning,* 353–360. ACM.

———. 2014. "Theory of disagreement-based active learning." *Foundations and Trends® in Machine Learning* 7 (2-3): 131–309.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. "Deep residual learning for image recognition." arXiv: `1512.03385`.

Heess, Nicolas, Gregory Wayne, David Silver, Tim Lillicrap, Tom Erez, and Yuval Tassa. 2015. "Learning Continuous Control Policies by Stochastic Value Gradients." In *Advances in Neural Information Processing Systems 28,* edited by C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, 2944–2952. Curran Associates, Inc.

Hibbard, Bill. 2012. "Model-based utility functions." *Journal of Artificial General Intelligence* 3 (1): 1–24.

Hinton, Geoffrey E, and Ruslan R Salakhutdinov. 2006. "Reducing the dimensionality of data with neural networks." *Science* 313 (5786): 504–507.

Hinton, Geoffrey, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. 2012. "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups." *IEEE Signal Processing Magazine* 29 (6): 82–97.

Huang, Ling, Anthony D. Joseph, Blaine Nelson, Benjamin I.P. Rubinstein, and J. D. Tygar. 2011. "Adversarial Machine Learning." In *4th ACM Workshop on Security and Artificial Intelligence,* 43–58. AISec '11. Chicago, Illinois, USA: ACM.

Hutter, Marcus. 2005. *Universal Artificial Intelligence: Sequential Decisions Based On Algorithmic Probability.* Texts in Theoretical Computer Science. Berlin: Springer.

Janzing, Dominik, David Balduzzi, Moritz Grosse-Wentrup, Bernhard Schölkopf, et al. 2013. "Quantifying causal influences." *The Annals of Statistics* 41 (5): 2324–2358.

Judah, Kshitij, Alan P. Fern, Thomas G. Dietterich, and Prasad Tadepalli. 2014. "Active Imitation Learning: Formal and Practical Reductions to I.I.D. Learning." *Journal of Machine Learning Research* 15:4105–4143.

Karpathy, Andrej, and Li Fei-Fei. 2015. "Deep Visual-Semantic Alignments for Generating Image Descriptions." In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* June.

Khani, Fereshte, and Martin Rinard. 2016. "Unanimous Prediction for 100% Precision with Application to Learning Semantic Mappings." *arXiv preprint arXiv:1606.06368.*

Kingma, Diederik P, and Max Welling. 2013. "Auto-encoding variational bayes." arXiv: 1312.6114 [cs.LG].

Klyubin, Alexander S, Daniel Polani, and Chrystopher L Nehaniv. 2005. "Empowerment: A universal agent-centric measure of control." In *Evolutionary Computation, 2005. The 2005 IEEE Congress on,* 1:128–135. IEEE.

Knox, W Bradley, and Peter Stone. 2009. "Interactively shaping agents via human reinforcement: The TAMER framework." In *Proceedings of the fifth international conference on Knowledge capture,* 9–16. ACM.

Korattikara, Anoop, Vivek Rathod, Kevin Murphy, and Max Welling. 2015. "Bayesian Dark Knowledge." arXiv: 1506.04416 [cs.LG].

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey Hinton. 2012. "Imagenet classification with deep convolutional neural networks." In *Advances in neural information processing systems,* 1097–1105.

Lake, Brenden M, Ruslan Salakhutdinov, and Joshua B Tenenbaum. 2015. "Human-level concept learning through probabilistic program induction." *Science* 350 (6266): 1332–1338.

Legg, Shane, and Marcus Hutter. 2007. "Universal Intelligence: A Definition of Machine Intelligence." *Minds and Machines* 17 (4): 391–444.

Letham, Benjamin, Cynthia Rudin, Tyler McCormick, David Madigan, et al. 2015. "Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model." *The Annals of Applied Statistics* 9 (3): 1350–1371.

Li, Guangliang, Shimon Whiteson, W Bradley Knox, and Hayley Hung. 2015. "Using informative behavior to increase engagement while learning from human reward." *Autonomous Agents and Multi-Agent Systems:* 1–23.

Li, Lihong, Michael L. Littman, and Thomas J. Walsh. 2008. "Knows What It Knows: A Framework for Self-aware Learning." In *25th International Conference on Machine Learning,* 568–575. ICML '08. Helsinki, Finland: ACM.

Liu, Huan, and Hiroshi Motoda. 2007. *Computational methods of feature selection.* CRC Press.

Mahendran, Aravindh, and Andrea Vedaldi. 2015. "Understanding deep image representations by inverting them." In *2015 IEEE conference on computer vision and pattern recognition (CVPR),* 5188–5196. IEEE.

Mnih, Volodymyr, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. 2013. "Playing Atari with Deep Reinforcement Learning." In *Deep Learning Workshop at Neural Information Processing Systems 26 (NIPS 2013).* ArXiv:1312.5602 [cs.LG]. Lake Tahoe, NV, USA.

Mnih, Volodymyr, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, et al. 2016. "Human-Level Control through Deep Reinforcement Learning." *Nature* 518, no. 7540 (February 26): 529–533.

Murphy, Tom. 2013. "The first level of Super Mario Bros. is easy with lexicographic orderings and time travel." *SIGBOVIK.*

Ng, Andrew Y., and Stuart J. Russell. 2000. "Algorithms for Inverse Reinforcement Learning." In *17th International Conference on Machine Learning (ICML-'00),* edited by Pat Langley, 663–670. San Francisco: Morgan Kaufmann.

Nguyen, Anh, Jason Yosinski, and Jeff Clune. 2015. "Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images." In *Computer Vision and Pattern Recognition, 2015 IEEE Conference on,* 427–436. IEEE.

Omohundro, Stephen M. 2008. "The Basic AI Drives." In *Artificial General Intelligence 2008: 1st AGI Conference,* edited by Pei Wang, Ben Goertzel, and Stan Franklin, 483–492. Frontiers in Artificial Intelligence and Applications 171. Amsterdam: IOS.

Orseau, Laurent, and Stuart Armstrong. 2016. "Safely Interruptible Agents." In *Uncertainty in Artificial Intelligence: 32nd Conference (UAI 2016),* edited by Alexander Ihler and Dominik Janzing, 557–566. Jersey City, New Jersey, USA.

Orseau, Laurent, and Mark Ring. 2011. "Self-Modification and Mortality in Artificial Agents." In Schmidhuber, Thórisson, and Looks 2011, 1–10.

Pearl, Judea. 2000. *Causality: Models, Reasoning, and Inference.* 1st ed. New York: Cambridge University Press.

———. 2009. *Causality: Models, Reasoning, and Inference.* 2nd ed. New York: Cambridge University Press.

Pulina, Luca, and Armando Tacchella. 2010. "An abstraction-refinement approach to verification of artificial neural networks." In *International Conference on Computer Aided Verification,* 243–257. Springer.

Rabin, Steve. 2010. *Introduction to game development.* Nelson Education.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier." *arXiv preprint arXiv:1602.04938.*

Ring, Mark, and Laurent Orseau. 2011. "Delusion, Survival, and Intelligent Agents." In *Artificial General Intelligence: 4th International Conference, (AGI 2011),* edited by Jürgen Schmidhuber, Kristinn R. Thórisson, and Moshe Looks, 11–20. Berlin, Heidelberg: Springer Berlin Heidelberg.

Robnik-Šikonja, Marko, and Igor Kononenko. 2008. "Explaining classifications for individual instances." *IEEE Transactions on Knowledge and Data Engineering* 20 (5): 589–600.

Rosenthal, Stephanie, Sai P. Selvaraj, and Manuela Veloso. 2016. "Verbalization: Narration of Autonomous Mobile Robot Experience." In *26th International Joint Conference on Artificial Intelligence (IJCAI'16).* New York City, NY.

Ross, Stéphane, Geoffrey J Gordon, and J Andrew Bagnell. 2010. "A reduction of imitation learning and structured prediction to no-regret online learning." arXiv: `1011.0686 [cs.LG]`.

Rubner, Yossi, Carlo Tomasi, and Leonidas J. Guibas. 2000. "The Earth Mover's Distance as a Metric for Image Retrieval." *International Journal of Computer Vision* 40 (2): 99–121.

Russell, Stuart J. 2014. "Of Myths and Moonshine." *Edge* (blog comment). `http://edge.org/conversation/the-myth-of-ai#26015`.

Russell, Stuart J., Daniel Dewey, and Max Tegmark. 2015. "Research Priorities for Robust and Beneficial Artificial Intelligence: An Open Letter." *AI Magazine* 36 (4).

Russell, Stuart J., and Peter Norvig. 2010. *Artificial Intelligence: A Modern Approach.* 3rd ed. Upper Saddle River, NJ: Prentice-Hall.

Salge, Christoph, Cornelius Glackin, and Daniel Polani. 2014. "Empowerment–an introduction." In *Guided Self-Organization: Inception,* 67–114. Springer.

Schmidhuber, Jürgen, Kristinn R. Thórisson, and Moshe Looks, eds. 2011. *Artificial General Intelligence: 4th International Conference, AGI 2011.* Lecture Notes in Computer Science 6830. Berlin: Springer.

Settles, Burr. 2010. "Active learning literature survey." *University of Wisconsin, Madison* 52 (55-66): 11.

Seung, H Sebastian, Manfred Opper, and Haim Sompolinsky. 1992. "Query by committee." In *5th annual workshop on Computational Learning Theory,* 287–294. ACM.

Siddiqui, Md Amran, Alan Fern, Thomas G. Dietterich, and Shubhomoy Das. 2016. "Finite Sample Complexity of Rare Pattern Anomaly Detection." In *Uncertainty in Artificial Intelligence: Proceedings of the 32nd Conference (UAI-2016),* edited by Alexander Ihler and Dominik Janzing, 686–695. Corvallis, Oregon: AUAI Press.

Silver, David, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. "Mastering the game of Go with deep neural networks and tree search." *Nature* 529 (7587): 484–489.

Simon, Herbert A. 1956. "Rational Choice and the Structure of the Environment." *Psychological Review* 63 (2): 129–138.

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. 2013. "Deep inside convolutional networks: Visualising image classification models and saliency maps." *arXiv preprint arXiv:1312.6034.*

Soares, Nate. 2016. "The Value Learning Problem." In *Ethics for Artificial Intelligence Workshop at IJCAI-16.* New York, NY.

Soares, Nate, and Benja Fallenstein. 2014. *Agent Foundations for Aligning Machine Intelligence with Human Interests: A Technical Research Agenda.* Technical report 2014—8. Forthcoming 2017 in "The Technological Singularity: Managing the Journey" Jim Miller, Roman Yampolskiy, Stuart J. Armstrong, and Vic Callaghan, Eds. Berkeley, CA: Machine Intelligence Research Institute.

Soares, Nate, Benja Fallenstein, Eliezer Yudkowsky, and Stuart Armstrong. 2015. "Corrigibility." In *1st International Workshop on AI and Ethics at AAAI-2015.* Austin, TX.

Štrumbelj, Erik, and Igor Kononenko. 2014. "Explaining prediction models and individual predictions with feature contributions." *Knowledge and information systems* 41 (3): 647–665.

Stuhlmüller, Andreas, Jessica Taylor, and Noah Goodman. 2013. "Learning stochastic inverses." In *Advances in neural information processing systems,* 3048–3056.

Szita, István, and Csaba Szepesvári. 2011. "Agnostic KWIK learning and efficient approximate reinforcement learning." In *JMLR.*

Taylor, Jessica. 2015. "Quantilizers: A Safer Alternative to Maximizers for Limited Optimization." In *2nd International Workshop on AI, Ethics and Society at AAAI-2016.* Phoenix, AZ.

Thomaz, Andrea L, and Cynthia Breazeal. 2006. "Transparency and socially guided machine learning." In *5th Intl. Conf. on Development and Learning (ICDL).*

Vellido, Alfredo, José David Martín-Guerrero, and Paulo Lisboa. 2012. "Making machine learning models interpretable." In *ESANN,* 12:163–172. Citeseer.

Von Neumann, John, and Oskar Morgenstern. 1944. *Theory of Games and Economic Behavior.* 1st ed. Princeton, NJ: Princeton University Press.

Vovk, Vladimir, Alex Gammerman, and Glenn Shafer. 2005. *Algorithmic learning in a random world.* Springer Science & Business Media.

Weber, Philippe, Gabriela Medina-Oliva, Christophe Simon, and Benoit Iung. 2012. "Overview on Bayesian networks applications for dependability, risk analysis and maintenance areas." *Engineering Applications of Artificial Intelligence* 25 (4): 671–682.

Wiskott, Laurenz, and Terrence J. Sejnowski. 2002. "Slow Feature Analysis: Unsupervised Learning of Invariances." *Neural Comput.* (Cambridge, MA, USA) 14, no. 4 (April): 715–770. ISSN: 0899-7667. doi:10.1162/089976602317318938. http://dx.doi.org/10.1162/089976602317318938.

Xu, Kelvin, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention." arXiv: 1502.03044 [cs.LG].

Yao, Yuan, Lorenzo Rosasco, and Andrea Caponnetto. 2007. "On early stopping in gradient descent learning." *Constructive Approximation* 26 (2): 289–315.

Yudkowsky, Eliezer. 2008. "Artificial Intelligence as a Positive and Negative Factor in Global Risk." In *Global Catastrophic Risks,* edited by Nick Bostrom and Milan M. Ćirković, 308–345. New York: Oxford University Press.

Zeiler, Matthew D, and Rob Fergus. 2014. "Visualizing and understanding convolutional networks." In *European Conference on Computer Vision,* 818–833. Springer.

Ziebart, Brian D, Andrew L Maas, J Andrew Bagnell, and Anind K Dey. 2008. "Maximum Entropy Inverse Reinforcement Learning." In *AAAI,* 1433–1438.