Aligning Superintelligence with Human Interests: A Technical Research Agenda

Nate Soares and **Benja Fallenstein**
Machine Intelligence Research Institute
{nate,benja}@intelligence.org

Contents

1 Introduction

The characteristic that has enabled humanity to shape the world is not strength, not speed, but intelligence. Barring catastrophe, it seems clear that progress in AI will one day lead to the creation of agents meeting or exceeding human-level general intelligence, and this will likely lead to the eventual development of systems which are "superintelligent" in the sense of being "smarter than the best human brains in practically every field" (Bostrom 2014). A superintelligent system could have an enormous impact upon humanity: just as human intelligence has allowed the development of tools and strategies that let humans control the environment to an unprecedented degree, a superintelligent system would likely be capable of developing tools and strategies that give it extraordinary power (Muehlhauser and Salamon 2012). In light of this potential, it is essential to use caution when developing artificially intelligent systems capable of attaining or creating superintelligence.

There is no reason to expect artificial agents to be driven by human motivations such as lust for power, but almost all goals can be better met with more resources (Omohundro 2008). This suggests that, by default, superintelligent agents would have incentives to acquire resources currently being used by humanity. (Can't we share? Likely not: there is no reason to expect artificial agents to be driven by human motivations such as fairness, compassion, or conservatism.) Thus, most goals would put the agent at odds with human interests, giving it incentives to deceive or manipulate its human operators and resist interventions designed to change or debug its behavior (Bostrom 2014, chap. 8).

Care must be taken to avoid constructing systems that exhibit this default behavior. In order to ensure that the development of smarter-than-human intelligence has a positive impact on humanity, we must

meet three formidable challenges: How can we create an agent that will reliably pursue the goals it is given? How can we formally specify beneficial goals? And how can we ensure that this agent will assist and cooperate with its programmers as they improve its design, given that mistakes in the initial version are inevitable?

This agenda discusses technical research that is tractable today, which the authors think will make it easier to confront these three challenges in the future. Sections 2 through 4 motivate and discuss six research topics that we think are relevant to these challenges. Section 5 discusses our reasons for selecting these six areas in particular.

We call a smarter-than-human system that reliably pursues beneficial goals "aligned with human interests" or simply "aligned." To become confident that an agent is aligned in this way, a practical implementation that merely *seems* to meet the challenges outlined above will not suffice. It is also necessary to gain a solid theoretical understanding of why that confidence is justified. This technical agenda argues that there is foundational research approachable today that will make it easier to develop aligned systems in the future, and describes ongoing work on some of these problems.

Of the three challenges, the one giving rise to the largest number of currently tractable research questions is the challenge of finding an agent architecture that will reliably pursue the goals it is given—that is, an architecture which is alignable in the first place. This requires theoretical knowledge of how to design agents which reason well and behave as intended even in situations never envisioned by the programmers. The problem of highly reliable agent designs is discussed in Section 2.

The challenge of developing agent designs which are tolerant of human error has also yielded a number of tractable problems. We argue that smarter-than-human systems would by default have incentives to manipulate and deceive the human operators. Therefore, special care must be taken to develop agent architectures which avert these incentives and are otherwise tolerant of programmer error. This problem and some related open questions are discussed in Section 3.

Reliable, error-tolerant agent designs are only beneficial if they are aligned with human interests. The difficulty of concretely specifying what is meant by "beneficial behavior" implies a need for some way to construct agents that reliably *learn* what to value (Bostrom 2014, chap. 12). A solution to this "value learning" problem is vital; attempts to start making progress are reviewed in Section 4.

Why these problems? Why now? Section 5 answers these questions and others. In short, the authors believe that there is theoretical research which can be done today that will make it easier to design aligned smarter-than-human systems in the future.

2 Highly Reliable Agent Designs

Bird and Layzell (2002) describe a genetic algorithm which, tasked with making an oscillator, re-purposed the printed circuit board tracks on the motherboard as a makeshift radio to amplify oscillating signals from nearby computers. The algorithm would not have found this solution if simulated on a virtual circuit board possessing only the features that *seemed* relevant to the problem. Intelligent search processes in the real world have the ability to use resources in unexpected ways, e.g. by finding "shortcuts" or "cheats" not accounted for in a simplified model.

When constructing intelligent systems which learn and interact with all the complexities of reality, it is not sufficient to verify that the algorithm behaves well in test settings. Additional work is necessary to verify that the system will continue working as intended in application. This is especially true of systems possessing general intelligence at or above the human level: superintelligent machines might find strategies and execute plans beyond both the experience and imagination of the programmers, making the clever oscillator of Bird and Layzell look trite.

Smarter-than-human systems could have an enormous impact upon humanity (Bostrom 2014). With great potential may come great risk: a system that is not aligned with human interests could cause catastrophic damage (Yudkowsky 2008). Because the stakes are so high, testing combined with a gut-level intuition that the system will continue to work outside the test environment is insufficient, even if the testing is extensive. It is necessary to also have a *formal* understanding of precisely why the system is expected to behave well in application.

What constitutes a formal understanding? It seems essential to us to have both (1) an understanding of precisely what problem the system is intended to solve; and (2) an understanding of precisely why *this* practical system is expected to solve *that* abstract problem. The latter must wait for the development of practical smarter-than-human systems, but the former is a theoretical research problem that can be approached today.

A full description of the problem would reveal the conceptual tools needed to understand why practical heuristics are expected to work. By analogy, consider the game of chess. Before designing practical chess algorithms, it is necessary to possess not only a predicate describing checkmate, but also a description of the problem in term of trees and backtracking algorithms: Trees and backtracking do not immediately yield a practical solution—building a full game tree is infeasible—but they are the conceptual tools of computer chess. It would be quite difficult to justify confidence in a chess heuristic before understanding trees and backtracking.

While these conceptual tools may seem obvious in hindsight, they were not clear to foresight. Consider the famous essay by Edgar Allen Poe about Maelzel's Mechanical Turk (Poe 1836). It is in many ways re-

markably sophisticated: Poe compares the Turk to "the calculating machine of Mr. Babbage" and then remarks on how the two systems cannot be of the same kind, since in Babbage's algebraical problems each step follows of necessity, and so can be represented by mechanical gears making deterministic motions; while in a chess game, no move follows with necessity from the position of the board, and even if our own move followed with necessity, the opponent's would not. And so (argues Poe) we can see that chess cannot possibly be played by mere mechanisms, only by thinking beings. From Poe's state of knowledge, Shannon's (1950) description of an idealized solution in terms of backtracking and trees constitutes a great insight.

We must put theoretical foundations under the field of general intelligence, in the same sense that Shannon put theoretical foundations under the field of computer chess.

Won't the foundations be developed over time, during the normal course of AI research? This is possible: in the past, theory has often preceded application. But the converse is also true: in many cases, application has preceded theory. The claim of this technical agenda is that, in safety-critical applications where mistakes can put lives at risk, it is crucial that the theory come first.

A smarter-than-human agent would be embedded within and computed by a complex universe, learning about its environment and bringing about desirable states of affairs. How is this formalized? What metric captures the question of how well an agent would perform in the real world?[1]

Not all parts of the problem must be solved in advance: the task of designing smarter, safer, more reliable systems could be delegated to early smarter-than-human systems, if the research done by those early systems can be sufficiently trusted. It is important, then, to focus research efforts particularly on parts of the problem where an increased understanding is necessary to construct a minimal reliable generally intelligent system. Moreover, it is important to focus on aspects which are currently tractable, so that progress can in fact be made today, and on issues relevant to alignment in particular, which would not otherwise be studied over the course of "normal" AI research.

In this section, we discuss four candidate topics meeting these criteria: (1) *realistic world models,* the study of agents learning and pursuing goals while embedded within a physical world; (2) *decision theory,* the study of idealized decision-making procedures; (3) *logical uncertainty,* the study of reliable reasoning with bounded deductive capabilities; and (4) *Vingean reflection,* the study of reliable methods for reasoning about agents that are more intelligent than the reasoner. We will now discuss each of these topics in turn.

1. Legg and Hutter (2007) provide a preliminary answer to this question, by defining a "universal measure of intelligence" which scores how well an agent can learn the features of an external environment and maximize a reward function. This is the type of formalization we are looking for: a scoring metric which describes how well an agent would achieve some set of goals. However, while the Legg-Hutter metric is insightful, it makes a number of simplifying assumptions, and many difficult open questions remain (Soares 2015a).

2.1 Realistic World Models

Formalizing the problem of computer intelligence may seem easy in theory: encode some set of preferences as a utility function, and evaluate the expected utility that would be obtained if the agent were implemented. However, this is not a full specification: What is the set of "possible realities" used to model the world? Against what distribution over world models is the agent evaluated? How is a given world model used to score an agent? To ensure that an agent would work well in reality, it is first necessary to formalize the problem faced by agents learning (and acting in) arbitrary environments.

Solomonoff (1964) made an early attempt to tackle these questions by specifying an "induction problem" in which an agent must construct world models and promote correct hypotheses based on the observation of an arbitrarily complex environment, in a manner reminiscent of scientific induction. In this problem, the agent and environment are separate. The agent gets to see one bit from the environment in each turn, and must predict the bits which follow.

Solomonoff's induction problem answers all three of the above questions in a simplified setting: The set of world models is any computable environment (e.g., any Turing machine). In reality, the simplest hypothesis that predicts the data is generally correct, so agents are evaluated against a simplicity distribution. Agents are scored according to their ability to predict their next observation. These answers were insightful, and led to the development of many useful tools, including algorithmic probability and Kolmogorov complexity.

However, Solomonoff's induction problem does not fully capture the problem faced by an agent learning about an environment while embedded *within* it, as a subprocess. It assumes that the agent and environment are separated, save only for the observation channel. What is the analog of Solomonoff induction for agents that are embedded within their environment?

This is the question of *naturalized induction* (Bensinger 2013). Unfortunately, the insights of Solomonoff do not apply in the naturalized setting. In Solomonoff's setting, where the agent and environment are separated, one can consider arbitrary Turing machines to be "possible environments." But when the agent is embedded in the environment, consideration must be restricted to environments which embed the agent. Given an algorithm, what is the set of environments which embed that algorithm? Given that set, what is the analogue of a simplicity prior which captures the fact that simpler hypotheses are more often correct?

Technical problem (Naturalized Induction)**.** *What, formally, is the induction problem faced by an intelligent agent embedded within and computed by its environment? What is the set of environments which embed the agent? What constitutes a simplicity prior over that set? How is the agent scored? For discussion, see Soares (2015a).*

Just as a formal description of Solomonoff induction answered the above three questions in the context of an agent learning an external environment, a formal description of naturalized induction may well yield answers to those questions in the context where agents are embedded in and computed by their environment.

Of course, the problem of computer intelligence is not simply an induction problem: the agent must also interact with the environment. Hutter (2000) extends Solomonoff's induction problem to an "interaction problem," in which an agent must both learn and act upon its environment. In each turn, the agent both observes one input and writes one output, and the output affects the behavior of the environment. In this problem, the agent is evaluated in terms of its ability to maximize a reward function specified in terms of inputs. While this model does not capture the difficulties faced by agents which are embedded within their environment, it does capture a large portion of the problem faced by agents interacting with arbitrarily complex environments. Indeed, the interaction problem (and AIXI [Hutter 2000], its solution) are the basis for the "universal measure of intelligence" developed by Legg and Hutter (2007).

However, even barring problems arising from the agent/environment separation, the Legg-Hutter metric does not fully characterize the problem of computer intelligence. It scores agents according to their ability to maximize a reward function specified in terms of observation. Agents scoring well by the Legg-Hutter metric are extremely effective at ensuring their observations optimize a reward function, but these high-scoring agents are likely to be the type that find clever ways to seize control of their observation channel rather than the type that identify and manipulate the features in the world that the reward function was intended to proxy for (Soares 2015a). Reinforcement learning techniques which punish the agent for attempting to take control would only incentivize the agent to deceive and mollify the programmers until it found a way to gain a decisive advantage (Bostrom 2014, chap. 8).

The Legg-Hutter metric does not characterize the question of how well an algorithm would perform if implemented in reality: to formalize that question, a scoring metric must evaluate the resulting environment histories, not just the agent's observations (Soares 2015a).

But human goals are not specified in terms of environment histories, either: they are specified in terms of high-level notions such as "money" or "flourishing humans." Leaving aside problems of philosophy, imagine rating a system according to how well it achieves a straightforward, concrete goal, such as by rating how much diamond is in an environment after the agent has acted on it, where "diamond" is specified concretely in terms of an atomic structure. Now the goals are specified in terms of atoms, and the environment histories are specified in terms of Turing machines paired with an interaction history. How is the environment history evaluated in terms of atoms? This is the *ontology identification* problem.

Technical problem (Ontology Identification)**.** *Given goals specified in some ontology and a world model, how can the ontology of the goals be identified in the world model? What types of world models are amenable to ontology identification? For a discussion, see Soares (2015a).*

To evaluate world models, the world models must be evaluated in terms of the ontology of the goals. This may be difficult in cases where the ontology of the goals does not match reality: it is one thing to locate atoms in a Turing machine using an atomic model of physics, but it is another thing altogether to locate atoms in a Turing machine modeling quantum physics. De Blanc (2011) further motivates the idea that explicit mechanisms are needed to deal with changes in the ontology of the system's world model.

Agents built to solve the wrong problem—such as optimizing their observations—may well be capable of attaining superintelligence, but it is unlikely that those agents could be aligned with human interests (Bostrom 2014, chap. 12). A better understanding of naturalized induction and ontology identification is needed to fully specify the problem that intelligent agents would face when pursuing human goals while embedded within reality, and this increased understanding could be a crucial tool when it comes to designing highly reliable agents.

2.2 Decision Theory

Smarter-than-human systems must be trusted to make good decisions, but what does it mean for a decision to be "good"? Formally, given a description of an environment and an agent embedded within, how is the "best available action" identified, with respect to some set of preferences? This is the question of decision theory.

The answer may seem trivial, at least in theory: simply iterate over the agent's available actions, evaluate what would happen if the agent took that action, and then return whichever action leads to the most expected utility. But this is not a full specification: How are the "available actions" identified? How is what "would happen" defined?

The difficulty is easiest to illustrate in a deterministic setting. Consider a deterministic agent embedded in a deterministic environment. There is exactly one action that the agent will take. Given a set of actions that it "could take," it is necessary to evaluate, for each

action, what would happen if the agent took that action. But the agent will not take most of those actions. How is the counterfactual environment constructed, in which a deterministic algorithm "does something" that, in the real environment, it doesn't do? Answering this question requires a theory of counterfactual reasoning, and counterfactual reasoning is not well understood.

Technical problem (Theory of Counterfactuals)**.** *What theory of counterfactual reasoning can be used to specify a procedure which always identifies the best action available to a given agent in a given environment, with respect to a given set of preferences? For discussion, see Soares and Fallenstein (2014).*

Decision theory has been studied extensively by philosophers. The study goes back to Pascal, and has been picked up in modern times by Lehmann (1950), Wald (1939), Jeffrey (1983), Joyce (1999), Lewis (1981), Pearl (2000), and many others. However, no satisfactory method of counterfactual reasoning yet answers this particular question. To give an example of why counterfactual reasoning can be difficult, consider a deterministic agent playing against a perfect copy of itself in the classic prisoner's dilemma (Rapoport and Chammah 1965). The opponent is guaranteed to do the same thing as the agent, but the agents are "causally separated," in that the action of one cannot physically affect the action of the other.

What is the counterfactual world in which the agent on the left cooperates? It is not sufficient to consider changing the action of the agent on the left while holding the action of the agent on the right constant, because while the two are causally disconnected, they are logically constrained to behave identically. Standard causal reasoning, which neglects these logical constraints, misidentifies "defection" as the best strategy available to each agent even when they know they have identical source codes (Lewis 1979).[2] Satisfactory counterfactual reasoning must respect these logical constraints, but how are logical constraints formalized and identified? It is fine to say that, in the counterfactual where the agent on the left cooperates, all identical copies of it also cooperate; but what counts as an identical copy? What if the right agent runs the same algorithm written in a different programming language? What if it only does the same thing 98% of the time?

These questions are pertinent in reality: practical agents must be able to identify good actions in settings where other actors base their actions on imperfect (but well-informed) predictions of what the agent will do. Identifying the best action available to an agent requires taking the non-causal logical constraints into account. A satisfactory formalization of counterfactual reasoning requires the ability to answer questions about how other deterministic algorithms behave in the counterfactual world where the agent's deterministic algorithm does something it doesn't. However, the evaluation of "logical counterfactuals" is not yet well understood.

Technical problem (Logical Counterfactuals)**.** *Consider a counterfactual in which a given deterministic decision procedure selects a different action from the one it selects in reality. What are the implications of this counterfactual on other algorithms? Can logical counterfactuals be formalized in a satisfactory way? A method for reasoning about logical counterfactuals seems necessary in order to formalize a more general theory of counterfactuals. For a discussion, see Soares and Fallenstein (2014).*

Unsatisfactory methods of counterfactual reasoning (such as the causal reasoning of Pearl (2000)) seem powerful enough to support smarter-than-human intelligent systems, but systems using those reasoning methods could systematically act in undesirable ways (even if otherwise aligned with human interests).

To construct practical heuristics that are known to make good decisions, even when acting beyond the oversight and control of humans, it is essential to understand what is meant by "good decisions." This requires a formulation which, given a description of an environment, an agent embedded in that environment, and some set of preferences, identifies the best action available to the agent. While modern methods of counterfactual reasoning do not yet allow for the specification of such a formula, recent research has pointed the way towards some promising paths for future research.

For example, Wei Dai's "updateless decision theory" (UDT) is a new take on decision theory that systematically outperforms causal decision theory (Hintze 2014), and two of the insights behind UDT highlight a number of tractable open problems (Soares and Fallenstein 2014).

Recently, Barasz et al. (2014) developed a concrete model, together with a Haskell implementation, of multi-agent games where agents have access to each others' source code and base their decisions on what they can prove about their opponent. They have found that it is possible for some agents to achieve robust cooperation in the one-shot prisoner's dilemma while remaining unexploitable (Barasz et al. 2014).

These results suggest a number of new ways to approach the problem of counterfactual reasoning, and we are optimistic that continued study will prove fruitful.

2.3 Logical Uncertainty

Consider a reasoner encountering a black box with one input chute and two output chutes. Inside the box is a complex Rube Goldberg machine that takes an input

2. As this is a multi-agent scenario, the problem of counterfactuals can also be thought of as game-theoretic. The goal is to define a procedure which reliably identifies the best available action; the label of "decision theory" is secondary. This goal subsumes both game theory and decision theory: the desired procedure must identify the best action in all settings, even when there is no clear demarcation between "agent" and "environment." Game theory informs, but does not define, this area of research.

ball and deposits it in one of the two output chutes. A probabilistic reasoner may have uncertainty about where the ball will exit, due to uncertainty about which Rube Goldberg machine is in the box. However, standard probability theory assumes that if the reasoner *did* know which machine the box implemented, they would know where the ball would exit: the reasoner is assumed to be *logically omniscient*, i.e., to know all logical consequences of any hypothesis they entertain.

By contrast, a practical bounded reasoner must be able to know exactly which Rube Goldberg machine the box implements without knowing where the ball will come out, due to the complexity of the machine. This reasoner is *logically uncertain*. Almost all practical reasoning is done under some form of logical uncertainty (Gaifman 2004), and almost all reasoning done by a smarter-than-human agent must be some form of logically uncertain reasoning. Any time an agent reasons about the consequences of a plan, the effects of running a piece of software, or the implications of an observation, it must do some sort of reasoning under logical uncertainty. Indeed, the problem of an agent reasoning about an environment in which it is embedded as a subprocess is inherently a problem of reasoning under logical uncertainty.

Thus, to construct a highly reliable smarter-than-human system, it is vitally important to ensure that the agent's logically uncertain reasoning is reliable and trustworthy. This requires a better understanding of the theoretical underpinnings of logical uncertainty, to more fully characterize what it means for logically uncertain reasoning to be "reliable and trustworthy" (Soares and Fallenstein 2015).

It is natural to consider extending standard probability theory to include the consideration of worlds which are "logically impossible" (e.g., where a deterministic Rube Goldberg machine behaves in a way that it doesn't). This gives rise to two questions: What, precisely, are logically impossible possibilities? And, given some means of reasoning about impossible possibilities, what is a reasonable prior probability distribution over impossible possibilities?

The problem is difficult to approach in full generality, but a study of logical uncertainty in the restricted context of assigning probabilities to logical sentences goes back at least to Łoś (1955) and Gaifman (1964), and has been investigated by many, including Halpern (2003), Hutter et al. (2013), Demski (2012), Russell (2014), and others. Though it isn't clear to what degree this formalism captures the kind of logically uncertain reasoning a realistic agent would use, logical sentences in, for example, the language of Peano Arithmetic are quite expressive: for example, given the Rube Goldberg machine discussed above, it is possible to form a sentence which is true if and only if the machine deposits the ball into the top chute. Thus, considering reasoners which are uncertain about logical sentences is a useful starting point. The problem of assigning probabilities to sentences of logic naturally divides itself into two parts.

First, how can probabilities consistently be assigned to sentences? An agent assigning probability 1 to short contradictions is hardly reasoning about the sentences as if they are logical sentences: some of the logical structure must be preserved. But which aspects of the logical structure? Preserving all logical implications requires that the reasoner be deductively omnipotent, as some implications $\phi \to \psi$ may be very involved. The standard answer in the literature is that a coherent assignment of probabilities to sentences corresponds to a probability distribution over complete, consistent logical theories (Gaifman 1964; Christiano 2014a); that is, an "impossible possibility" is any consistent assignment of truth values to all sentences. Deductively limited reasoners cannot have fully coherent distributions, but they can approximate these distributions: for a deductively limited reasoner, "impossible possibilities" can be any assignment of truth values to sentences that looks consistent so far, so long as the assignment is discarded as soon as a contradiction is introduced.

Technical problem (Impossible Possibilities)**.** *How can deductively limited reasoners approximate reasoning according to a probability distribution over complete theories of logic? For a discussion, see Christiano (2014a).*

Second, what is a satisfactory prior probability distribution over logical sentences? If the system is intended to reason according to a theory at least as powerful as Peano Arithmetic (PA), then that theory will be incomplete (Gödel, Kleene, and Rosser 1934). What prior distribution places nonzero probability on the set of complete extensions of PA? Deductively limited agents would not be able to literally use such a prior, but if it were computably approximable, then they could start with a rough approximation of the prior and refine it over time. Indeed, the process of refining a logical prior—getting better and better probability estimates for given logical sentences—captures the whole problem of reasoning under logical uncertainty in miniature. Hutter et al. (2013) has defined a desirable prior, but Sawin and Demski (2013) have shown that it cannot be computably approximated. Demski (2012) and Christiano (2014a) have also proposed logical priors, but neither seems fully satisfactory. The specification of satisfactory logical priors is difficult in part because it is not yet clear which properties are desirable in a logical prior, nor which properties are possible.

Technical problem (Logical Priors)**.** *What is a satisfactory set of priors over logical sentences that a bounded reasoner can approximate? For a discussion, see Soares and Fallenstein (2015).*

Many existing tools for studying reasoning, such as game theory, standard probability theory, and Bayesian networks, all assume that reasoners are logically omniscient. A theory of reasoning under logical uncertainty seems necessary to formalize the problem of naturalized

induction, and to generate a satisfactory theory of counterfactual reasoning. If these tools are to be developed, extended, or improved, then a better understanding of logically uncertain reasoning is required.

2.4 Vingean Reflection

Instead of specifying superintelligent systems directly, it seems likely that humans may instead specify generally intelligent systems that go on to create or attain superintelligence. In this case, the reliability of the resulting superintelligent system depends upon the reasoning of the initial system which created it (either anew or via self-modification).

If the agent reasons reliably under logical uncertainty, then it may have a generic ability to evaluate various plans and strategies, only selecting those which seem beneficial. However, some scenarios put that logically uncertain reasoning to the test more than others. There is a qualitative difference between reasoning about simple programs and reasoning about human-level intelligent systems. For example, modern program verification techniques could be used to ensure that a "smart" military drone obeys certain rules of engagement, but it would be a different problem altogether to verify the behavior of an artificial military general which must run an entire war. A general has far more autonomy, ability to come up with clever unexpected strategies, and opportunities to impact the future than a drone.

A self-modifying agent (or any agent which constructs new agents more intelligent than itself) must reason about the behavior of an agent which is more intelligent than the reasoner. This type of reasoning is critically important to the design of self-improving agents: if a system will attain self-modification through superintelligence, then the impact of the system depends entirely upon the correctness of the original agent's reasoning about its self-modifications (Fallenstein and Soares 2015).

Before trusting a system to attain superintelligence, it seems prudent to ensure that the agent uses appropriate caution when reasoning about successor agents.[3] That is, it seems necessary to understand the mechanisms by which agents reason about smarter systems.

Naive tools for reasoning about plans including smarter agents, such as backwards induction (Ben-Porath 1997), would have the reasoner evaluate the smarter agent by simply checking what the smarter agent would do. This does not capture the difficulty of the problem: a parent agent cannot simply check what its successor agent would do in all scenarios, for if it could, then it would already know what actions its successor will take, and the successor would not in any way be smarter.

Yudkowsky and Herreshoff (2013) call this observation the "Vingean principle," after Vernor Vinge (1993), who emphasized how difficult it is for humans to predict the behavior of smarter-than-human agents. Any agent reasoning about more intelligent successor agents must do so *abstractly*, without pre-computing all actions that the successor would take in every scenario. We refer to this kind of reasoning as *Vingean reflection*.

Technical problem (Vingean Reflection)**.** *How can agents reliably reason about agents which are smarter than themselves, without violating the Vingean principle? For discussion, see Fallenstein and Soares (2015).*

It may seem premature to worry about how agents reason about self-improvements before developing a theoretical understanding of reasoning under logical uncertainty in general. However, it seems to us that work in this area can inform understanding of what sort of logically uncertain reasoning is necessary to reliably handle Vingean reflection.

Given the high stakes when constructing systems smarter than themselves, artificial agents might use some form of extremely high-confidence reasoning to verify the safety of potentially dangerous self-modifications. When *humans* desire extremely high reliability, as is the case for (e.g.) autopilot software, we often use formal logical systems (US DoD 1985; UK MoD 1991). High-confidence reasoning in critical situations may require something akin to formal verification (even if *most* reasoning is done using more generic logically uncertain reasoning), and so studying Vingean reflection in the domain of formal logic seems like a good starting point.

Logical models of agents reasoning about agents that are "more intelligent," however, run into a number of obstacles. By Gödel's second incompleteness theorem (1934), sufficiently powerful formal systems cannot rule out the possibility that they may be inconsistent. This makes it difficult for agents using formal logical reasoning to verify the reasoning of similar agents which also use formal logic for high-confidence reasoning; the first agent cannot verify that the latter agent is consistent. Roughly, it seems desirable to be able to develop agents which reason as follows:

> This smarter successor agent uses reasoning similar to mine, and my own reasoning is sound, so its reasoning is sound as well.

However, Gödel, Kleene, and Rosser (1934) showed that this sort of reasoning leads to inconsistency, and these problems do in fact make Vingean reflection difficult in a logical setting (Fallenstein and Soares 2015; Yudkowsky 2013).

Technical problem (Löbian Obstacle)**.** *How can agents gain very high confidence in agents that use similar reasoning systems, while avoiding paradoxes of self-*

3. Of course, if an agent reasons *perfectly* under logical uncertainty, it would also reason well about the construction of successor agents. However, given the fallibility of human reasoning and the fact that this path is critically important, it seems prudent to verify the agent's reasoning methods in this scenario specifically.

reference? For discussion, see Fallenstein and Soares (2015).

These results may seem like artifacts of models rooted in formal logic, and may seem irrelevant given that practical agents must eventually use logical uncertainty rather than formal logic to reason about smarter successors. However, it has been shown that many of the Gödelian obstacles carry over into early probabilistic logics in a straightforward way, and some results have already been shown to apply in the domain of logical uncertainty (Fallenstein 2014).

Study into toy models of the formal logical setting has led to partial solutions (Fallenstein and Soares 2014). Recent work has pushed these models towards probabilistic settings (Fallenstein and Soares 2014; Yudkowsky 2014; Soares 2014). Further research may continue driving the development of methods for reasoning under logical uncertainty which can handle Vingean reflection in a reliable way (Fallenstein and Soares 2015).

3 Error-Tolerant Agent Designs

Incorrectly specified superintelligent agents could be dangerous (Yudkowsky 2008). Correcting a modern AI system involves simply shutting the system down and modifying its source code. Modifying a smarter-than-human system may prove more difficult: systems attaining superintelligence could acquire new hardware, alter its software, create subagents, and take other actions that would leave the original programmers with only dubious control over the agent. This is especially true if the agent has incentives to resist modification or shutdown. If intelligent systems are to be safe, they must be constructed in such a way that they are amenable to correction, even if they have the ability to prevent or avoid correction.

This does not come for free: by default, agents have incentives to preserve their own preferences, even if those conflict with the intentions of the programmers (Omohundro 2008; Soares et al., forthcoming). Special care is needed to specify agents that avoid the default incentives to manipulate and deceive (Bostrom 2014, chap. 8).

Restricting the actions available to a superintelligent agent may be quite difficult (chap. 9). Intelligent optimization processes often find unexpected ways to fulfill their optimization criterion using whatever resources are at their disposal; recall the evolved oscillator of Bird and Layzell (2002) which re-purposed printed circuit tracks as a makeshift radio. Superintelligent optimization processes may well use hardware, software, and other resources in unanticipated ways, making them difficult to contain if they have incentives to escape.

We must learn how to design agents which do not have incentives to escape, manipulate, or deceive in the first place: agents which reason as if they are incomplete and potentially flawed in dangerous ways, and which are therefore amenable to online correction. Reasoning of this form is known as "corrigible reasoning."

Technical problem (Corrigibility)**.** *What sort of reasoning can reflect the fact that an agent is incomplete and potentially flawed in dangerous ways? For discussion, see Soares et al. (forthcoming).*

Naïve attempts at specifying corrigible behavior are unsatisfactory: For example, "moral uncertainty" frameworks could allow agents to learn values through observation and interaction, but would still incentivize agents to resist changes to the underlying moral uncertainty framework if it happened to be flawed. Simple "penalty terms" for manipulation and deception also seem doomed to failure: such agents would have incentives to resist modification while cleverly avoiding the technical definitions of "manipulation" and "deception." The goal is not to design systems that fail in their attempts to deceive the programmers; the goal is to understand reasoning methods that do not give rise to deception incentives in the first place.

A formalization of the intuitive notion of corrigibility remains elusive. Active research is currently focused on small toy problems, in the hopes that insight gained there will generalize. One such toy problem is the "shutdown problem," which involves designing a set of preferences that incentivize an agent to shutdown upon the press of a button without also incentivizing the agent to either cause or prevent the pressing of that button (Soares et al., forthcoming). Stuart Armstrong's utility indifference technique (forthcoming) provides a partial solution, but not a fully satisfactory one.

Technical problem (Utility Indifference)**.** *Can a utility function be specified such that agents maximizing that utility function switch their preferences on demand, without having incentives to cause or prevent the switching? For discussion, see Armstrong (forthcoming).*

A better understanding of corrigible reasoning is essential to design agent architectures that are tolerant of human error. Other research could also prove fruitful, including research into reliable containment mechanisms. Alternatively, agent designs could somehow incentivize the agent to have a "low impact" on the world. Specifying "low impact" is trickier than it may seem: How do you tell an agent that it can't affect the physical world, given that its RAM is physical? How do you tell it that it can only use its own hardware, without allowing it to use its motherboard as a makeshift radio? How do you tell it not to cause big changes in the world when its behavior influences the actions of the programmers, who influence the world in chaotic ways?

Technical problem (Domesticity)**.** *How can an intelligent agent be safely incentivized to have a low impact? Specifying such a thing is not as easy as it seems. For a discussion, see Armstrong, Sandberg, and Bostrom (2012).*

Regardless of the methodology used, it is crucial to understand designs for agents that could be updated and modified during the development process, so as to ensure that the inevitable human errors do not lead to catastrophe.

4 The Value Learning Problem

A highly-reliable, error-tolerant agent design does not guarantee positive impact: the benefit of the system still depends entirely upon whether it is given appropriate goals.

A superintelligent system may find clever, unintended ways to achieve the specific goals that it is given. Imagine a superintelligent system designed to cure cancer which does so by stealing resources, proliferating robotic laboratories at the expense of the biosphere, and kidnapping test subjects: the intended goal may have been "cure cancer without doing anything bad," but such a goal is rooted in cultural context and shared human knowledge.

It is not sufficient to construct systems that are smart enough to figure out the intended goals: Human beings, upon learning that natural selection "intended" sex to be pleasurable only for purposes of reproduction, do not suddenly decide that contraceptives are abhorrent. While one should not anthropomorphize natural selection, humans are capable of understanding the process which created them while being completely unmotivated to alter their preferences. For similar reasons, when developing artificial intelligent agents, is not sufficient to develop a system intelligent enough to figure out the intended goals; the system must also be explicitly constructed to pursue them (Bostrom 2014, chap. 8).

However, the "intentions" of the operators are a complex, vague, fuzzy, context-dependent notion (Yudkowsky 2011). Concretely writing out the full intentions of the operators in a machine-readable format is implausible if not impossible, even for simple tasks. An intelligent agent must be designed to learn and act according to the preferences of its operators.[4] This is the *value learning problem*.

Directly programming a rule which identifies cats in images is implausibly difficult, but specifying a system which inductively learns how to identify cats in images is possible. Similarly, while directly programming a rule capturing complex human intentions is implausibly difficult, intelligent agents could be constructed to inductively learn values from training data.

Inductive value learning presents unique difficulties. The goal is to develop a system which can classify potential outcomes according to their value, but what sort of training data allows this classification? The labeled data could be given in terms of the agent's world model, but this is a brittle solution if the ontology of the world model is liable to change. Alternatively, the labeled data could come in terms of observations, in which case the agent would have to first learn how the labels in the observations map onto objects in the world model, and *then* learn how to classify outcomes. Designing algorithms which can do this likely requires a better understanding of methods for constructing multi-level world models from sense data.

Technical problem (Multi-level World Models)**.** *How can multi-level world models be constructed from sense data in a manner amenable to ontology identification? For a discussion, see Soares (2015b).*

Standard problems of inductive learning arise, as well: how could a training set be constructed which allows the agent to fully learn the complexities of value? It is easy to imagine a training set which labels many observations of happy humans as "good" and many observations of needlessly suffering humans as "bad," but the simplest generalization from this data set may well be that humans value human-shaped things mimicking happy emotions: after training on this data, an agent may be inclined to construct many simple animatronics mimicking happiness. Creating a training set that covers all relevant dimensions of human value is difficult for the same reason that specifying human value directly is difficult. In order for inductive value learning to succeed, it is necessary to construct a system which identifies ambiguities in the training set—dimensions along which the training set gives no information—and queries the operators accordingly.

Technical problem (Ambiguity Identification)**.** *Given a training data set and a world model, how can dimensions which are neglected by the training data be identified? For discussion, see Soares (2015b).*

This problem is not unique to value learning, but it is especially important for it. Research into the programmatic identification of ambiguities, and the generation of "queries" which are similar to previous training data but differ along the ambiguous axis, would assist in the development of systems which can safely perform inductive value learning.

Intuitively, an intelligent agent should be able to use some of its intelligence to assist in this process: it does not take a detailed understanding of the human psyche to deduce that humans care more about some ambiguities (are the human-shaped things actually humans?) than others (does it matter if there is a breeze?). To build a system that acts as intended, the system must model the intentions of the operators and act accordingly. This adds another layer of indirection: the system must model the operators in some way, and must extract "preferences" from the operator-model and update its preferences accordingly (in a manner robust

4. Or of all humans, or of all sapient creatures, etc. There are many philosophical concerns surrounding what sort of goals are ethical when aligning a superintelligent system, but a solution to the value learning problem is necessary regardless.

against improvements to the model of the operator). Techniques such as inverse reinforcement learning (Ng and Russell 2000), in which the agent assumes that the operator is maximizing some reward function specified in terms of observations, are a good start, but many questions remain unanswered.

Technical problem (Operator Modeling)**.** *By what methods can an operator be modeled in such a way that (1) a model of the operator's preferences can be extracted; and (2) the model may eventually become arbitrarily accurate and represent the operator as a subsystem embedded within the larger world? For a discussion, see Soares (2015b).*

A system which acts as the operators intend may still have significant difficulty answering questions that the operators themselves cannot answer: imagine humans trying to design an artificial agent to do what they would want, if they were better people. How can normative uncertainty (uncertainty about moral claims) be resolved? Bostrom (2014, chap. 13) suggests an additional level of indirection: task the system with reasoning about what sorts of conclusions the operators would come to if they had more information and more time to think. Formalizing this is difficult, and the problems are largely still in the realm of philosophy rather than technical research. However, Christiano (2014b) has sketched one possible method by which the volition of a human could be extrapolated, and Soares (2015b) discusses some potential pitfalls.

Philosophical problem (Normative Uncertainty)**.** *What ought one do when one is uncertain about what one ought to do? What norms govern uncertainty about normative claims? For a discussion, see MacAskill (2014).*

Human operators with total control over a superintelligent system could give rise to a moral hazard of extraordinary proportions by putting unprecedented power into the hands of a small few (Bostrom 2014, chap. 6). The extraordinary potential of superintelligence gives rise to many questions of morality and ethics. When constructing systems intended to have a great ability to control the future, it is important to consider the construction of agents which not only act according to the intentions of the operators, but also in the interest of all humanity or perhaps all sentient life. Here we largely leave the philosophical questions aside, and remark only that those who design systems intended for superintelligence will take on a responsibility of unprecedented scale.

5 Discussion

Sections 2 through 4 discussed six research topics where the authors think that further research could make it easier to develop aligned systems in the future. This section discusses our reasons for selecting these six topics in particular, and the reasons why we think that useful progress can be made today.

5.1 Research that Cannot be Delegated

Creating a superintelligent system aligned with human interests does not require specifying an aligned superintelligent system from scratch. On the path to great intelligence, much of the work may be done by smarter-than-human systems. Since the Dartmouth Proposal (McCarthy et al. 1955), it has been a standard idea in AI that a sufficiently smart machine intelligence could be intelligent enough to improve itself. In 1965, I. J. Good observed that this might create a positive feedback loop leading to an "intelligence explosion" (Good 1965). Bostrom (2014, chap. 4) has observed that an intelligence explosion is especially likely if the agent has the ability to acquire more hardware, improve its software, or design new hardware.

With this in mind, we must therefore focus on research that cannot be safely delegated to machines. Corrigibility research is a good example: the design of corrigible reasoning can hardly be delegated to agents that have incentives to manipulate and deceive. To construct corrigible systems, humans must first gain a better understanding of corrigibility.

By contrast, consider the fields of computer vision or natural language processing, where intelligent agents would have incentives to correct and improve the relevant tools so as to enhance their ability to model and interact with the world.

The topics discussed in this agenda are ones that we believe are difficult to delegate to intelligent agents. Why can't these tasks, too, be delegated? Why not, e.g., design a system that makes "good enough" decisions, constrain it to domains where its decisions are trusted, and then let it develop a better decision theory, perhaps using an indirect normativity approach (chap. 13) to figure out how humans would have wanted it to make decisions?

We cannot delegate these tasks because modern knowledge is not sufficient even for an indirect approach. Even if fully satisfactory theories of logical uncertainty and decision theory cannot be obtained, it is still necessary to have a sufficient theoretical grasp on the obstacles in order to justify high confidence in the system's ability to correctly perform indirect normativity.

Furthermore, it would be risky to delegate a crucial task before attaining a solid theoretical understanding of exactly what task is being delegated. It is possible to create an intelligent system tasked with developing better and better approximations of Bayesian updating, but it would be difficult to delegate the abstract task of "find good ways to update probabilities" to an intelligent system *before* gaining an understanding of Bayesian reasoning. The theoretical understanding is

necessary to ensure that the right questions are being asked.

5.2 Topics that are Tractable, Uncrowded, and Focused

This technical agenda primarily covers topics that we believe are *tractable*, that is, topics which contain concrete open problems, where progress could be made immediately. Ultimately, significant effort will be required to actually align real smarter-than-human systems with human interests, but in lieu of working designs for smarter-than-human systems, it is difficult if not impossible to begin that work in advance. Instead, this agenda focuses on research geared towards gaining a better understanding of the problems faced by an intelligent system embedded within reality. Regardless of whether practical smarter-than-human systems arise in ten years or in one hundred years, we will be better able to design safe systems if we understand these problems.

This agenda further limits attention to *uncrowded* domains, where there is not already an abundance of research being done, and where the problems may not be solved over the course of "normal" AI research. For example, program verification techniques are absolutely crucial in the design of extremely reliable programs, but program verification is not covered in this agenda primarily because a vibrant community is already actively studying the topic.

Finally, this agenda restricts consideration to topics that are focused on developing tools useful for designing aligned systems in particular (as opposed to intelligent systems in general). It may be possible to develop a practical understanding of intelligence that is sufficient to spark an intelligence explosion before developing a theoretical understanding of reasoning that is sufficient to construct a reliably aligned system. This could result in scenarios where teams have incentives to cut corners or take risks, and these incentives could be dangerous (Bostrom 2014, chap. 14). Currently, significant research effort is focused on improving the capabilities of artificially intelligent systems, and comparatively little effort is focused on superintelligence alignment (Bostrom 2014, chap. 14). Therefore, this agenda focuses on research that improves the ability to design aligned systems in particular.

5.3 Theoretical Research Approachable Today

Why is so much of this technical agenda focused on topics such as decision theory and logical uncertainty? Shouldn't superintelligence alignment research focus primarily on AI constraint or on value learning? Some think that this agenda sounds more like generic theoretical AI research than alignment-specific research.

Progress on the topics outlined in this agenda could indeed make it easier to design intelligent systems in general. Just as the intelligence metric of Legg and Hutter (2007) lent insight into the ideal priors for agents facing Hutter's interaction problem, a full description of the naturalized induction problem could lend insight into the ideal priors for agents embedded within their universe. A satisfactory theory of logical uncertainty could lend insight into general intelligence more broadly. An ideal decision theory could reveal an ideal decision-making procedure for real agents to approximate.

But while these advancements might provide tools useful for designing intelligent systems in general, they would make it drastically easier to design aligned systems in particular. Idealized solutions, while impractical, provide the conceptual tools necessary to reason about practical solutions. Though no chess-playing program evaluates a full game tree, it would be very difficult to design a reliable chess program without first understanding the conceptual tools of backtracking algorithms and search trees (or something equivalent). It is much easier to design trustworthy heuristics after figuring out exactly what solution the heuristic is supposed to approximate.

Conversely, if we must evaluate real systems composed of practical heuristics before formalizing the theoretical problems that those heuristics are supposed to solve, then we will be forced to rely on unreliable intuition; we will be vulnerable to bias and overconfidence while attempting to provide an answer before we have finished understanding the question.

The theoretical understanding might not be developed by default. Causal counterfactual reasoning, despite being unsatisfactory, may be good enough to enable the construction of a smarter-than-human system. Heuristics for reasoning under logical uncertainty could yield smarter-than-human systems that work for reasons we don't quite understand. Systems built from unsatisfactory and poorly understood heuristics might be capable of creating or attaining superintelligence—but it is unlikely that such systems could be aligned with human interests.

Sometimes theory precedes application, but sometimes it does not. The goal of much of the research outlined in this agenda is to ensure, in the domain of superintelligence alignment—where the stakes are incredibly high—that theoretical understanding comes first.

5.4 What If This Work is Irrelevant?

This is a hazard of any attempt to do work in advance. The question is not whether the work has a *certainty* of being helpful, the question is whether it's wiser to start now or to try to come up with all the relevant theory at the last minute. We have identified a number of unanswered foundational questions relating to the development of general intelligence, and at present it seems possible to make promising progress. The prudent course, then, is to begin as soon as possible.

5.5 Why Start Now?

It is prudent to develop a theory of superintelligence alignment before developing a system capable of attaining or creating superintelligence. It may seem premature to tackle the problem now, with superintelligent systems still firmly in the domain of futurism. But imagine the chagrin if, in a few decades, the need for a mature theory of corrigibility is imminent, but the field is just as immature as seen in this technical agenda!

We think it is wise to approach these problems as soon as they look approachable. To do otherwise seems to us like a cognitive bias surrounding the fear of wasted effort, rather than a prudent calculation of the probable consequences of doing something versus nothing.

Weld and Etzioni (1994) made a call to arms, noting that "society will reject autonomous agents unless we have some credible means of making them safe." We are concerned with the opposite problem: what if society fails to reject systems that are unsafe? What will be the consequences, if someone believes a smarter-than-human system is aligned with human interests, when it is not?

This document is our call to arms: regardless of whether or not research efforts follow the path laid out in this document, significant research effort must be focused on the study of superintelligence alignment as soon as possible.

References

Armstrong, Stuart. Forthcoming. "AI Motivated Value Selection." Accepted to the 1st International Workshop on AI and Ethics, held within the 29th AAAI Conference on Artificial Intelligence (AAAI-2015), Austin, TX.

Armstrong, Stuart, Anders Sandberg, and Nick Bostrom. 2012. "Thinking Inside the Box: Controlling and Using an Oracle AI." *Minds and Machines* 22 (4): 299–324. doi:10.1007/s11023-012-9282-2.

Barasz, Mihaly, Paul Christiano, Benja Fallenstein, Marcello Herreshoff, Patrick LaVictoire, and Eliezer Yudkowsky. 2014. "Robust Cooperation in the Prisoner's Dilemma: Program Equilibrium via Provability Logic." Unpublished manuscript, January 23. http://arxiv.org/pdf/1401.5577v1.pdf.

Ben-Porath, Elchanan. 1997. "Rationality, Nash Equilibrium, and Backwards Induction in Perfect-Information Games." *Review of Economic Studies* 64 (1): 23–46. doi:10.2307/2971739.

Bensinger, Rob. 2013. "Building Phenomenological Bridges." *Less Wrong* (blog), December 23. http://lesswrong.com/lw/jd9/building_phenomenological_bridges/.

Bird, Jon, and Paul Layzell. 2002. "The Evolved Radio and Its Implications for Modelling the Evolution of Novel Sensors." In *Proceedings of the 2002 Congress on Evolutionary Computation,* 2:1836–1841. Honolulu, HI: IEEE. doi:10.1109/CEC.2002.1004522.

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

Christiano, Paul. 2014a. *Non-Omniscience, Probabilistic Inference, and Metamathematics.* Machine Intelligence Research Institute, Berkeley, CA, June 22. http://intelligence.org/files/Non-Omniscience.pdf.

———. 2014b. "Specifying 'enlightened judgment' precisely (reprise)." *Ordinary Ideas* (blog), August 27. http://ordinaryideas.wordpress.com/2014/08/27/specifying-enlightened-judgment-precisely-reprise/.

De Blanc, Peter. 2011. *Ontological Crises in Artificial Agents' Value Systems.* The Singularity Institute, San Francisco, CA, May 19. http://arxiv.org/abs/1105.3821.

Demski, Abram. 2012. "Logical Prior Probability." In *Artificial General Intelligence: 5th International Conference, AGI 2012, Oxford, UK, December 8–11, 2012. Proceedings,* edited by Joscha Bach, Ben Goertzel, and Matthew Iklé, 50–59. Lecture Notes in Artificial Intelligence 7716. New York: Springer. doi:10.1007/978-3-642-35506-6_6.

Fallenstein, Benja. 2014. *Procrastination in Probabilistic Logic.* Machine Intelligence Research Institute, Berkeley, CA. http://intelligence.org/files/ProbabilisticLogicProcrastinates.pdf.

Fallenstein, Benja, and Nate Soares. 2014. "Problems of Self-Reference in Self-Improving Space-Time Embedded Intelligence." In *Artificial General Intelligence: 7th International Conference, AGI 2014, Quebec City, QC, Canada, August 1–4, 2014. Proceedings,* edited by Ben Goertzel, Laurent Orseau, and Javier Snaider, 21–32. Lecture Notes in Artificial Intelligence 8598. New York: Springer. doi:10.1007/978-3-319-09274-4_3.

———. 2015. *Vingean Reflection: Reliable Reasoning for Self-Improving Agents.* Machine Intelligence Research Institute, Berkeley, CA.

Gaifman, Haim. 1964. "Concerning Measures in First Order Calculi." *Israel Journal of Mathematics* 2 (1): 1–18. doi:10.1007/BF02759729.

———. 2004. "Reasoning with Limited Resources and Assigning Probabilities to Arithmetical Statements." *Synthese* 140 (1–2): 97–119. doi:10.1023/B:SYNT.0000029944.99888.a7.

Gödel, Kurt, Stephen Cole Kleene, and John Barkley Rosser. 1934. *On Undecidable Propositions of Formal Mathematical Systems.* Princeton, NJ: Institute for Advanced Study.

Good, Irving John. 1965. "Speculations Concerning the First Ultraintelligent Machine." In *Advances in Computers,* edited by Franz L. Alt and Morris Rubinoff, 6:31–88. New York: Academic Press. doi:10.1016/S0065-2458(08)60418-0.

Halpern, Joseph Y. 2003. *Reasoning about Uncertainty.* Cambridge, MA: MIT Press.

Hintze, Daniel. 2014. *Problem Class Dominance in Predictive Dilemmas.* Machine Intelligence Research Institute, Berkeley, CA, April 23. http://intelligence.org/files/ProblemClassDominance.pdf.

Hutter, Marcus. 2000. "A Theory of Universal Artificial Intelligence based on Algorithmic Complexity." Unpublished manuscript, April 3. http://arxiv.org/abs/cs/0004001.

Hutter, Marcus, John W. Lloyd, Kee Siong Ng, and William T. B. Uther. 2013. "Probabilities on Sentences in an Expressive Logic." *Journal of Applied Logic* 11 (4): 386–420. doi:10.1016/j.jal.2013.03.003.

Jeffrey, Richard C. 1983. *The Logic of Decision.* 2nd ed. Chicago: Chicago University Press.

Joyce, James M. 1999. *The Foundations of Causal Decision Theory.* Cambridge Studies in Probability, Induction and Decision Theory. New York: Cambridge University Press. doi:10.1017/CBO9780511498497.

Legg, Shane, and Marcus Hutter. 2007. "Universal Intelligence: A Definition of Machine Intelligence." *Minds and Machines* 17 (4): 391–444. doi:10.1007/s11023-007-9079-x.

Lehmann, E. L. 1950. "Some Principles of the Theory of Testing Hypotheses." *Annals of Mathematical Statistics* 21 (1): 1–26. doi:10.1214/aoms/1177729884.

Lewis, David. 1979. "Prisoners' Dilemma is a Newcomb Problem." *Philosophy & Public Affairs* 8 (3): 235–240. http://www.jstor.org/stable/2265034.

———. 1981. "Causal Decision Theory." *Australasian Journal of Philosophy* 59 (1): 5–30. doi:10.1080/00048408112340011.

Łoś, Jerzy. 1955. "On the Axiomatic Treatment of Probability." *Colloquium Mathematicae* 3 (2): 125–137. http://eudml.org/doc/209996.

MacAskill, William. 2014. "Normative Uncertainty." PhD diss., St Anne's College, University of Oxford. http://ora.ox.ac.uk/objects/uuid:8a8b60af-47cd-4abc-9d29-400136c89c0f.

McCarthy, John, Marvin Minsky, Nathan Rochester, and Claude Shannon. 1955. *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence.* Stanford, CA: Formal Reasoning Group, Stanford University, August 31.

Muehlhauser, Luke, and Anna Salamon. 2012. "Intelligence Explosion: Evidence and Import." In *Singularity Hypotheses: A Scientific and Philosophical Assessment,* edited by Amnon Eden, Johnny Søraker, James H. Moor, and Eric Steinhart. The Frontiers Collection. Berlin: Springer.

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

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

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

Poe, Edgar Allan. 1836. "Maelzel's Chess-Player." *Southern Literary Messenger* 2 (5): 318–326.

Rapoport, Anatol, and Albert M. Chammah. 1965. *Prisoner's Dilemma: A Study in Conflict and Cooperation.* Vol. 165. Ann Arbor Paperbacks. Ann Arbor: University of Michigan Press.

Russell, Stuart. 2014. "Unifying Logic and Probability: A New Dawn for AI?" In *Information Processing and Management of Uncertainty in Knowledge-Based Systems: 15th International Conference, IPMU 2014, Montpellier, France, July 15-19, 2014, Proceedings, Part I,* 10–14. Communications in Computer and Information Science 442. Springer. doi:10.1007/978-3-319-08795-5_2.

Sawin, Will, and Abram Demski. 2013. *Computable probability distributions which converge on* Π_1 *will disbelieve true* Π_2 *sentences.* Machine Intelligence Research Institute, Berkeley, CA, July. http://intelligence.org/files/Pi1Pi2Problem.pdf.

Shannon, Claude E. 1950. "XXII. Programming a Computer for Playing Chess." *Philosophical Magazine,* Series 7, 41 (314): 256–275. doi:10.1080/14786445008521796.

Soares, Nate. 2014. *Tiling Agents in Causal Graphs.* Machine Intelligence Research Institute, Berkeley, CA, May. http://intelligence.org/files/TilingAgentsCausalGraphs.pdf.

———. 2015a. *Formalizing Two Problems of Realistic World-Models.* Machine Intelligence Research Institute, Berkeley, CA.

———. 2015b. *The Value Learning Problem.* Machine Intelligence Research Institute, Berkeley, CA.

Soares, Nate, and Benja Fallenstein. 2014. *Toward Idealized Decision Theory.* Machine Intelligence Research Institute, Berkeley, CA.

———. 2015. *Questions of Reasoning Under Logical Uncertainty.* Machine Intelligence Research Institute, Berkeley, CA.

Soares, Nate, Benja Fallenstein, Eliezer Yudkowsky, and Stuart Armstrong. Forthcoming. "Corrigibility." Accepted to the 1st International Workshop on AI and Ethics, held within the 29th AAAI Conference on Artificial Intelligence (AAAI-2015), Austin, TX.

Solomonoff, Ray J. 1964. "A Formal Theory of Inductive Inference. Part I." *Information and Control* 7 (1): 1–22. doi:10.1016/S0019-9958(64)90223-2.

United Kingdom Ministry of Defense. 1991. *Requirements for the Procurement of Safety Critical Software in Defence Equipment.* Interim Defence Standard 00-55. United Kingdom Ministry of Defense, April 5.

United States Department of Defense. 1985. *Department of Defense Trusted Computer System Evaluation Criteria.* Department of Defense Standard DOD 5200.28-STD. United States Department of Defense, December 26. http://csrc.nist.gov/publications/history/dod85.pdf.

Vinge, Vernor. 1993. "The Coming Technological Singularity: How to Survive in the Post-Human Era." In *Vision-21: Interdisciplinary Science and Engineering in the Era of Cyberspace,* 11–22. NASA Conference Publication 10129. NASA Lewis Research Center. http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19940022855_1994022855.pdf.

Wald, Abraham. 1939. "Contributions to the Theory of Statistical Estimation and Testing Hypotheses." *Annals of Mathematical Statistics* 10 (4): 299–326. doi:10.1214/aoms/1177732144.

Weld, Daniel, and Oren Etzioni. 1994. "The First Law of Robotics (A Call to Arms)." In *Proceedings of the Twelfth National Conference on Artificial Intelligence,* edited by Barbara Hayes-Roth and Richard E. Korf, 1042–1047. Menlo Park, CA: AAAI Press. http://www.aaai.org/Papers/AAAI/1994/AAAI94-160.pdf.

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.

———. 2011. "Complex Value Systems in Friendly AI." In *Artificial General Intelligence: 4th International Conference, AGI 2011, Mountain View, CA, USA, August 3–6, 2011. Proceedings,* edited by Jürgen Schmidhuber, Kristinn R. Thórisson, and Moshe Looks, 388–393. Lecture Notes in Computer Science 6830. Berlin: Springer. doi:10.1007/978-3-642-22887-2_48.

———. 2013. *The Procrastination Paradox.* Brief Technical Note. Machine Intelligence Research Institute, Berkeley, CA. http://intelligence.org/files/ProcrastinationParadox.pdf.

———. 2014. *Distributions Allowing Tiling of Staged Subjective EU Maximizers.* Machine Intelligence Research Institute, Berkeley, CA, May 11. Revised May 31, 2014. http://intelligence.org/files/DistributionsAllowingTiling.pdf.

Yudkowsky, Eliezer, and Marcello Herreshoff. 2013. *Tiling Agents for Self-Modifying AI, and the Löbian Obstacle.* Early Draft. Machine Intelligence Research Institute, Berkeley, CA. http://intelligence.org/files/TilingAgents.pdf.