Aligning Superintelligence with Human Interests: An Annotated Bibliography

Nate Soares
Machine Intelligence Research Institute
nate@intelligence.org

Abstract

How could superintelligent systems be aligned with the interests of humanity? This annotated bibliography compiles some recent research relevant to that question, and categorizes it into six topics: (1) realistic world models; (2) idealized decision theory; (3) logical uncertainty; (4) Vingean reflection; (5) corrigibility; and (6) value learning. Within each subject area, references are organized in an order amenable to learning the topic. These are by no means the only six topics relevant to the study of alignment, but this annotated bibliography could be used by anyone who wants to understand the state of the art in one of these six particular areas of active research.

Background

Soares, Nate, and Benja Fallenstein. 2014. *Aligning Superintelligence with Human Interests: A Technical Research Agenda.* Technical report 2014–8. Berkeley, CA: Machine Intelligence Research Institute. `https://intelligence.org/files/TechnicalAgenda.pdf`.

The problem faced by an artificially intelligent agent acting in the real world is not yet well understood: even given unlimited computing resources, modern knowledge would not allow for the specification of an artificially intelligent system reliably aligned with human interests. This technical agenda argues that it is possible to do theoretical research today to put foundations under the field of alignment, laying the groundwork for the specification of safe systems in the future, and introduces a number of research areas. Why these topics? Why now? The document answers these questions and others, motivating the six areas of active research covered below.

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

Bostrom's *Superintelligence* is the canonical introduction to the topic of superintelligence alignment. Will superintelligence be developed? What kind, and when? Why is caution necessary? The arguments put forth in this text are greatly helpful to understanding the alignment research outlined in this bibliography.

Research supported by the Machine Intelligence Research Institute (intelligence.org). Published as Technical report 2015–5.

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.

Can we predict the behavior of a system which is smarter than us? While we would not be able to guess the specific plans of a more intelligent system, Omohundro argues that behaviors such as resource acquisition and self improvement are incentivized by almost any set of preferences. This implies that it is possible to reason about the behavior of superintelligent systems, today. (For an updated discussion of the same topic, see chapter chapter 7 of Bostrom’s *Superintelligence.*)

Realistic World Models

Soares, Nate. 2015. *Formalizing Two Problems of Realistic World-Models.* Technical report 2015–3. Berkeley, CA: Machine Intelligence Research Institute. `https://intelligence.org/files/RealisticWorldModels.pdf`.

Superintelligent systems must be embedded as a subprocess in the real world, and must reason about an environment which is larger than the system in order to achieve complex goals. Can this problem be fully formalized, in the same way that Hutter formalized the problem of agents interacting with an external environment? Soares and Fallenstein motivate the study of agents constructing realistic world models while embedded (and computed by) in a complex environment.

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

Hutter constructs a framework for studying agents which learn and act upon an external environment. This formalization lends some insight into the problem faced by intelligent systems acting in an arbitrarily complex environment. The formalism lends itself to the specification of AIXI, a solution which excels in Hutter’s model of interaction, and which lends some insight into the nature of intelligence.

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`.

Hutter’s interaction problem can be used to define a “measure of intelligence,” by evaluating how well the agent scores against a simplicity distribution over all possible computable environments. This constitutes an early attempt to fully describe the problem of agents interacting with an arbitrarily complex environment, by giving a metric by which scores intelligence in this setting.

Orseau, Laurent, and Mark Ring. 2012. “Space-Time Embedded Intelligence.” 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, 209–218. Lecture Notes in Artificial Intelligence 7716. New York: Springer. doi:`10.1007/978-3-642-35506-6_22`.

Orseau and Ring provide an alternative framework for evaluating intelligence, which allows for the fact that agents are not separate from their environment, but rather embedded in it and computed by it. Their formulation characterizes the problem which we, as the designers of AI systems, face in choosing the most effective system to implement.

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

Bensinger broaches the question of how to construct agents that score well on “space-time embedded” metrics: what sort of scientific induction could an agent perform in order to learn about the environment which embeds the agent as a subprocess? Attempts to formalize this question have encountered many difficulties.

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.

Given goals specified in terms of one ontology, how are these evaluated in terms of a world model written in a different ontology? Say that the goal is to produce diamonds, as specified by an atomic structure. How is this evaluated in terms of a world model representing a quantum universe? De Blanc discusses this problem, and notes that good behavior cannot be expected in intelligent systems by default.

Idealized Decision Theory

Soares, Nate, and Benja Fallenstein. 2014. *Toward Idealized Decision Theory.* Technical report 2014–7. Berkeley, CA: Machine Intelligence Research Institute. https://intelligence.org/files/TowardIdealizedDecisionTheory.pdf.

It is important that superintelligent systems tend to select good decisions, but what is a "good decision"? Can we give a definition which precisely describes how to identify the best action available to a given agent in a given situation (with respect to some set of preferences)? Intuitively, the problem may seem easy: simply identify the action which maximizes expected utility. However, this is not a full specification: if a deterministic agent is embedded in a deterministic environment, how is expected utility evaluated in the counterfactual case where a deterministic agent does something that it doesn't? This "counterfactual reasoning" is difficult to formalize, and no satisfactory theory of counterfactual reasoning yet exists. Soares and Fallenstein motivate the need for a better theory of decision-making before constructing smarter-than-human systems.

LaVictoire, Patrick, Benja Fallenstein, Eliezer Yudkowsky, Mihaly Barasz, Paul Christiano, and Marcello Herreshoff. 2014. "Program Equilibrium in the Prisoner's Dilemma via Löb's Theorem." In *Multiagent Interaction without Prior Coordination: Papers from the AAAI-14 Workshop.* AAAI Publications. http://www.aaai.org/ocs/index.php/WS/AAAIW14/paper/view/8833.

LaVictoire et al. have developed a mechanism by which agents can achieve robust mutual cooperation in the one-shot Prisoner's Dilemma (with shared source code), while remaining unexploitable. Their "modal agent" framework may provide a mechanism through which satisfactory methods of counterfactual reasoning (allowing e.g. unexploitable mutually cooperative behavior) can be studied.

Dai, Wei. 2009. "Towards a New Decision Theory." *Less Wrong* (blog) (August 13). http://lesswrong.com/lw/15m/towards_a_new_decision_theory/.

Recent progress in the development of a satisfactory theory of counterfactual reasoning has stemmed from the development of Dai's "updateless decision theory" (UDT), introduced in this internet blog post.

Fallenstein, Benja. 2012. "A Model of UDT with a Concrete Prior over Logical Statements." *Less Wrong* (blog) (August 28). http://lesswrong.com/lw/eaa/a_model_of_udt_with_a_concrete_prior_over_logical/.

In this blog post, Fallenstein gives a simple model of UDT which acts in a probabilistic setting.

Altair, Alex. 2013. *A Comparison of Decision Algorithms on Newcomblike Problems.* Machine Intelligence Research Institute, Berkeley, CA. http://intelligence.org/files/Comparison.pdf.

Altair compares several different approaches to decision theory, providing an overview of some pros and cons of each one.

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

Hintze demonstrates that Dai's updateless approach to decision theory dominates many other modern approaches, with respect to a certain class of decision problems.

Slepnev, Vladimir. 2011. "Example Decision Theory Problem: 'Agent simulates predictor.'" *Less Wrong* (blog) (May 19). http://lesswrong.com/lw/5rq/example_decision_theory_problem_agent_simulates/.

Despite encouraging results demonstrated by Hintze and others, UDT is not without flaws. This blog post by Slepnev introduces the Agent Simulates Predictor problem, which remains an open problem for UDT.

Benson-Tilsen, Tsvi. 2014. *UDT with Known Search Order.* Technical report 2014–4. Berkeley, CA: Machine Intelligence Research Institute. http://intelligence.org/files/UDTSearchOrder.pdf.

In this technical report, Benson-Tilsen describes "spurious proofs" which can lead proof-based models of UDT to perform suboptimally, along with a solution to this problem.

Fallenstein, Benja. 2014. "An Optimality Result for Modal UDT." http://forum.intelligence.org/item?id=50.

Despite its flaws, UDT is a very powerful decision theory. Fallenstein demonstrates that UDT is in fact optimal on a certain class of "fair" decision problems (with some caveats, of course).

Logical Uncertainty

Soares, Nate, and Benja Fallenstein. 2015. *Questions of Reasoning Under Logical Uncertainty.* Technical report 2015–1. Berkeley, CA: Machine Intelligence Research Institute. https://intelligence.org/files/QuestionsLogicalUncertainty.pdf.

How can agents reason as if they know the laws of logic, and the description of a computer program, but not the program's output? Despite significant study, a formal understanding of logically uncertain reasoning remains elusive. In this paper, Soares and Fallenstein motivate the need for a better theoretical understanding of reasoning under logical uncertainty before constructing smarter-than-human systems.

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

Gaifman introduces what is now the canonical answer to the question of how reasoners can consistently place probabilities on first-order logical sentences. (Answer: by assigning probabilities according to a distribution over complete theories of logic.)

———. 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.

Reasoning according to a measure over complete theories requires unlimited deductive capabilities. In this paper, Gaifman moves towards probability distributions over sentences that could be used with limited deduction.

Christiano, Paul F., Eliezer Yudkowsky, Marcello Herreshoff, and Mihaly Barasz. 2013. *Definability of Truth in Probabilistic Logic.* Working Paper. Machine Intelligence Research Institute, Berkeley, CA, April 2. https://intelligence.org/files/DefinabilityTruthDraft.pdf.

Christiano et al. demonstrate a "reflective" probabilistic logic, which assigns accurate probabilities to statements about its own probability assignments (up to infinitesimal error). This result demonstrated that probabilistic logics can avoid some of the paradoxes of self-reference common in traditional logic. (Unfortunately, this system was later shown to be unsound.)

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.

In this paper, Hutter describes a prior probability distribution over complete theories which has many desirable properties (but which is not computably approximable).

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`.

Demski proposes an alternative to Hutter’s prior which is computably approximable. Unfortunately, this prior lacks many nice properties.

Christiano, Paul. 2014. *Non-Omniscience, Probabilistic Inference, and Metamathematics.* Technical report 2014–3. Berkeley, CA: Machine Intelligence Research Institute. `http://intelligence.org/files/Non-Omniscience.pdf`.

Christiano gives a broad overview of many problems related to reasoning under logical uncertainty, and proposes a number of techniques which lend some new insight to the problem and point in the direction of practicality.

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`.

One of the problems with the development of satisfactory logical priors is that it is not clear what it means for a prior to be “satisfactory.” Sawin and Demski demonstrate that certain desirable properties cannot be possessed by computable prior probability distributions.

Hahn, Jeremy. 2013. *Scientific Induction in Probabilistic Mathematics.* Brief Technical Note. Machine Intelligence Research Institute, Berkeley, CA. `http://intelligence.org/files/ScientificInduction.pdf`.

Hahn reports on the difficulty of attaining another desirable property, which is roughly that priors which have seen that ϕ holds in 10% of cases should converge on probability 0.1 for new instances of ϕ. No known logical priors have this property.

Vingean Reflection

Fallenstein, Benja, and Nate Soares. 2015. *Vingean Reflection: Reliable Reasoning for Self-Improving Agents.* Technical report 2015–2. Berkeley, CA: Machine Intelligence Research Institute. `https://intelligence.org/files/VingeanReflection.pdf`.

In the case of smarter-than-human systems which self-modify (or otherwise create smarter agents), the behavior of the resulting system depends entirely upon the initial agent’s ability to reason about systems that are smarter than it. How can such reasoning be done reliably? In this paper, Fallenstein et al. motivate that further understanding of reliable reasoning about smarter agents is necessary before building any agent capable of significant reliable recursive self-improvement.

———. 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`.

Fallenstein and Soares provide a toy model of self-modifying agents which reason reliably about smarter agents using formal proofs in simplified settings.

Soares, Nate, and Benja Fallenstein. 2014. *Botworld.* Technical report 2014–2. Berkeley, CA: Machine Intelligence Research Institute. `http://intelligence.org/files/Botworld.pdf`.

The simplified setting above is a cellular automaton known as Botworld, introduced by Soares and Fallenstein in this technical report. Botworld serves as a concrete toy model in which it is easier to envision (and study) the behavior of self-modifying agents.

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`.

Yudkowsky and Herreshoff introduce a number of difficulties encountered in formal logical settings where agents attempt to reason about smarter agents. This paper introduces the "Löbian obstacle" to Vingean reflection, and discusses some partial solutions, such as Fallenstein's "parametric polymorphism".

Weaver, Nik. 2013. "Paradoxes of Rational Agency and Formal Systems That Verify Their Own Soundness." Unpublished manuscript, December 21. `http://arxiv.org/abs/1312.3626`.

Weaver gives a further account of the Löbian obstacle and its impact on agents using formal systems to reason about systems of similar capability.

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

Many systems which successfully avoid the "Löbian obstacle" are unsound, for reasons outlined by Yudkowsky in this technical report. High-confidence reasoning about smarter agents seems to require finding a fine balance between the Löbian obstacle and the procrastination paradox.

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

Unfortunately, the reflective system of probabilistic logic developed by Christiano et al. succumbs to the procrastination paradox, as demonstrated by Fallenstein in this technical report.

———. 2014b. *An Infinitely Descending Sequence of Sound Theories each Proving the Next Consistent.* Brief Technical Note. Machine Intelligence Research Institute, Berkeley, CA. `http://intelligence.org/files/ConsistencyWaterfall.pdf`.

This technical report by Fallenstein looks at toy models of agents attempting to reason about successors using proofs, and presents one of the most satisfactory partial solutions to date. This "consistency waterfall" avoids both the Löbian obstacle and the procrastination paradox (by placing certain constraints on the language in which goals can be expressed).

———. 2014c. *Decreasing Mathematical Strength in One Formalization of Parametric Polymorphism.* Brief Technical Note. Machine Intelligence Research Institute, Berkeley, CA. `http://intelligence.org/files/DecreasingStrength.pdf`.

Another partial solution to problems of self-reference in proof-based toy models is Fallenstein's "parametric polymorphism." In this technical report, Fallenstein describes one open problem relating to parametric polymorphism.

Yudkowsky, Eliezer. 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`.

In this technical report, Yudkowsky develops the proof-based toy models of Vingean reflection toward application in probabilistic domains.

Soares, Nate. 2014. *Tiling Agents in Causal Graphs.* Technical report 2014–5. Berkeley, CA: Machine Intelligence Research Institute. `http://intelligence.org/files/TilingAgentsCausalGraphs.pdf`.

Soares further develops the study of Vingean reflection into the domain of (probabilistic) causal graphs, by introducing four lemmas sufficient for proof-based agents to license the creation of agents of similar proof strength in a causal graph.

Corrigibility

Soares, Nate, and Benja Fallenstein. 2015. *Questions of Reasoning Under Logical Uncertainty.* Technical report 2015–1. Berkeley, CA: Machine Intelligence Research Institute. `https://intelligence.org/files/QuestionsLogicalUncertainty.pdf`.

By default, smarter-than-human systems would have incentives to manipulate and deceive programmers. How can these incentives be averted? What reasoning methods allow agents to reason as if they are under construction and potentially flawed in dangerous ways? Soares et al. introduce the field of "Corrigibility," which studies these questions. The paper goes on to discuss an early approach to one relevant subproblem.

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.

In this paper, Armstrong introduces a method by which an agent can be built to change its goals upon receiving a certain command, without also giving the agent incentives to cause or prevent such commands.

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`.

Can a very intelligent agent be made to only do a simple task, such as answer questions, without giving it other dangerous incentives as well? Maybe, but such a thing is not as easy as it may seem. Armstrong et al. discuss why.

Value Learning

Soares, Nate. 2015. *The Value Learning Problem.* Technical report 2015–4. Berkeley, CA: Machine Intelligence Research Institute. `https://intelligence.org/files/ValueLearningProblem.pdf`.

Human intentions are complex, vague, context-dependent, and culturally laden. A superintelligent machine would not, by default, act as intended. Soares and Yudkowsky discuss methods by which a system be constructed to reliably learn what to value and/or model the intentions of its operators and act accordingly.

Dewey, Daniel. 2011. "Learning What to Value." 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, 309–314. Lecture Notes in Computer Science 6830. Berlin: Springer. doi:`10.1007/978-3-642-22887-2_35`.

Dewey argues that reinforcement learning cannot lead to smarter-than-human agents that have a beneficial impact, and motivates the need for agents which learn what to value.

Yudkowsky, Eliezer. 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`.

In this paper, Yudkowsky further motivates the point that, in order to construct smarter-than-human systems aligned with human interests, it is necessary for the agent to learn an accurate representation of the complex human notion of "value.".

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

In *Superintelligence* (chapter 13), Bostrom describes "indirectly normative" approaches to value learning, such as Yudkowsky's "Coherent Extrapolated Volition." In this blog post, Christiano describes an alternative indirectly normative approach, by sketching how the "enlightened judgement" of an individual could be formalized using a specific counterfactual scenario.

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`.

Any agent using an indirectly normative approach must reason under "normative uncertainty," which is uncertainty about what sorts of things have moral value. In this dissertation, MacAskill explores the pros and cons of several different solutions (such as the "parliamentary model" suggested by Bostrom) in several different contexts.

Fallenstein, Benja, and Nisan Stiennon. 2014. *"Loudness": On Priors over Preference Relations.* Brief Technical Note. Machine Intelligence Research Institute, Berkeley, CA. `http://intelligence.org/files/LoudnessPriors.pdf`.

It is natural to consider toy models of "morally uncertain" agents by considering toy models of agents with uncertainty about their utility function. Such models run into trouble related to the fact that utility functions are equivalent under positive affine transformation.

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, CA.

Ng and Russell describe a method by which agents could learn the reward functions of other agents in the environment. This approach could provide an alternative method for value learning, but has yet to be fully explored in that context.