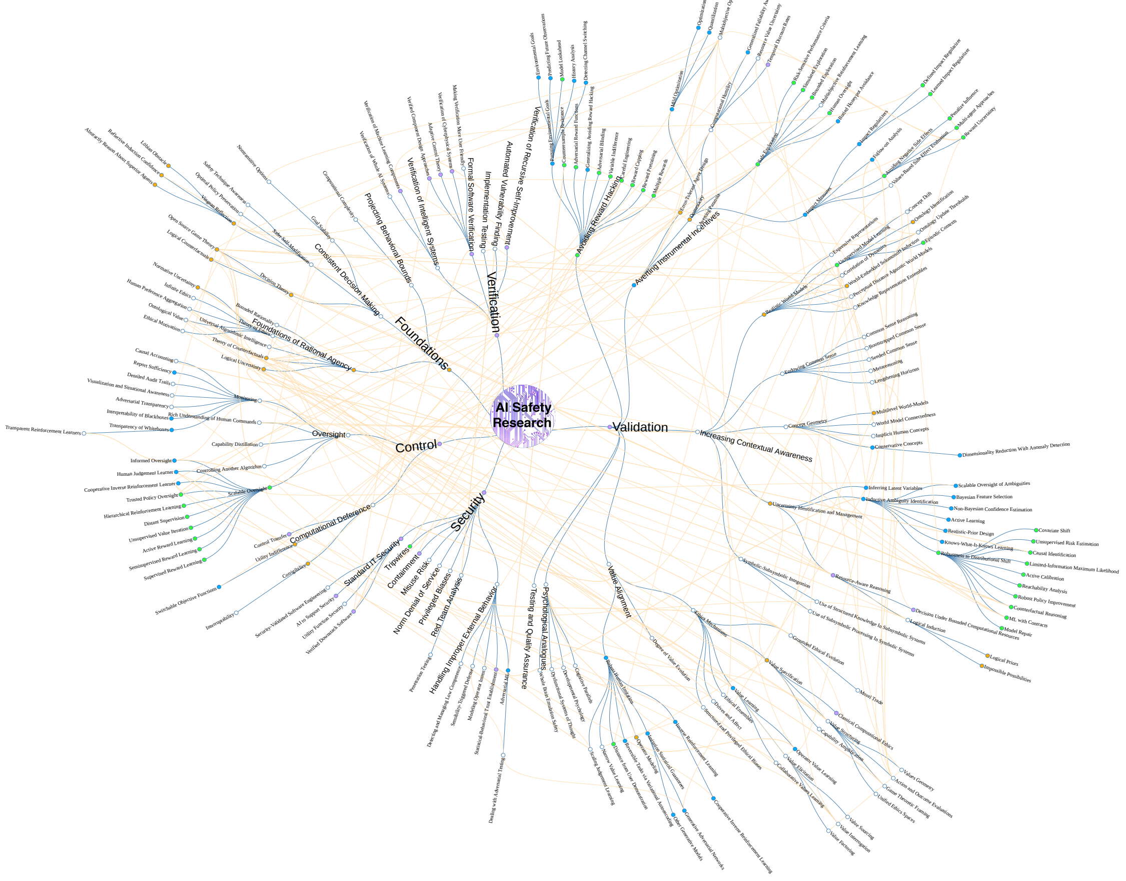
The Landscape of AI Safety and Beneficence Research: Input for Brainstorming at Beneficial AI 2017*

Executive Summary

Success in the quest for artificial intelligence has the potential to bring unprecedented benefits to humanity, and it is therefore worthwhile to research how to maximize these benefits while avoiding potential pitfalls. This document gives numerous examples of research topics aimed at ensuring that AI remains robust and beneficial.


AI Safety Research
Verification
Validation
Security
Control
Foundations
Foundations of Rational Agency
Consistent Decision Making
Projecting Behavioral Bounds
Verification of Intelligent Systems
Formal Software Verification
Implementation Testing
Automated Vulnerability Finding
Verification of Recursive Self-Improvement
Avoiding Reward Hacking
Averting Instrumental Incentives
Increasing Contextual Awareness
Value Alignment
Psychological Analogues
Testing and Quality Assurance
Handling Improper External Behavior
Red Team Analysis
Privileged Biases
Norm Denial of Service
Misuse Risk
Containment
Tripwires
Standard IT Security
Computational Deference
Oversight

*This document borrows heavily from MIRI's agent foundations technical agenda [403], Amodei et al.'s concrete problems review [11], MIRI's machine learning technical agenda [437], and FLI's previous research priorities document [371]. This document was written by Richard Mallah and was last edited January 2, 2017. It reflects feedback from Abram Demski, Andrew Critch, Anthony Aguirre, Bart Selman, Jan Leike, Jessica Taylor, Max Tegmark, Meia Chita-Tegmark, Nate Soares, Paul Christiano, Roman Yampolskiy, Stuart Armstrong, Tsvi Benson-Tilsen, Vika Krakovna, and Vincent Conitzer. (ver. 0.41)

Contents

1 Introduction

Surprising to many outside the field, within the field of artificial intelligence, asking "What if we succeed?" is actually only a rare afterthought [372, 373]. Many researchers believe that as machine intelligences become more general and broad in their capability [176], more capable in their powers, more powerful and reliable, and more heavily relied upon for important and complex roles, there is both significant opportunity for remarkable benefits and great peril in extraordinary risks [41]. More organizations are creating medium-term roadmaps to achieving advanced levels of AI [293, 357], and the advent of deep learning [173] has kickstarted a period of rapid developments in AI and ML [389, 287, 234]. As agents perform a wider variety of tasks [446], with more online learning occurring [390], and become more general-purpose [256, 234, 374, 43], challenges in safety, robustness, and beneficence increase substantially [368], and so this calls for increased focus, resources [229], and energy to be applied to AI safety.

It has been argued that an artificial intelligence smarter than the best minds in almost every field can of course have an enormous impact on humanity [58, 368]. Some researchers believe it's possible or even likely that an advanced agent might not be entirely aligned with the values of humanity or even its operators by default, and it would have tendencies to choose to continue its own existence, and to acquire resources, in order to achieve the goals it has been given [367], since nearly all goals can be better met with more time and more resources [321]. In these scenarios, such an agent would likely be capable of developing its own tools and strategies for exerting additional control [304]. Some scientists argue that since artificial agents would not share human evolutionary history, there is no reason to expect them to be driven by human motivations by default [475], and they would not *automatically* acquire a human sense of fairness, compassion, or conservatism [403]. A system attaining superintelligence, as defined [58], 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 [403]. It's argued that unpredictable behavior from highly capable agents can cause catastrophic damage if not aligned with human interests, and can bring great benefits if they are [478, 41, 130]. There has been discussion of how society should respond to such risks [58, 413], but broader society is only now realizing the potential for current work to significantly mitigate such risks. There is much relevant foundational research that can be worked on, today and within the next couple of decades, that will make it easier to develop aligned systems in the future [403].

This document broadly outlines the set of ongoing work in the very related endeavors of AI safety, robustness, and beneficence (sometimes all together just called AI safety for short). *It borrows heavily from MIRI's agent foundations technical agenda [403], Amodei et al.'s concrete problems review [11], MIRI's machine learning technical agenda [437], and FLI's previous research priorities document [371].* The field of AI safety, robustness, and beneficence aims to develop theoretical considerations for AI safety, to develop formal understandings of the problems, to create proofs placing limits on what can expected where possible, and to create practical techniques and algorithms that provide understood degrees of assurance. Recognized experts in the field have argued that as there are many teams working on both AI and AI safety under greatly different assumptions, it seems prudent to maintain a portfolio approach to the field at this time [371]. This field is rather young; the vast majority of techniques listed in this document are still in exploratory phases, and there are likely many as-yet-unknown topics and many kinds of unknown unknowns [303] regarding what will be most effective for the purpose. High-level goals of this research include that powerful AI be safely overseen by human operators, that it pursues objectives that humans actually care about, reflecting values that stakeholders approve of, and that it fails gracefully and humbly: creating systems that do what is *good* rather than just what is *powerful.* The purpose of this document is to provide a comprehensive map and reference set for the field.

There are many opportunities for AI researchers to expand upon this research, to add safety and beneficence requirements in their own problem definitions, and to explore intersections between their work and safety work. Some of these techniques can be applicable for nearer-term systems, and a subset of those might be expected to scale to extremely powerful and general systems [11]. Yet other near-term safety techniques can be useful for long-term safety research by pointing in promising directions or by teaching us about what will not scale. One advantage of near-term applicability in both of these cases is the ability to implement and test them [323] without some of the more challenging philosophical or foundational questions first needing to be addressed [120, 255]. Additional theoretical progress is quite necessary, however, in order to confidently understand what will be safe going forward and have assurances as to the reasons [403].

2 Foundations

There are a number of open foundational mathematical and philosophical problems that have bearing on multiple facets of safety [403].

2.1 Foundations of Rational Agency

Within foundational mathematics, logic, and philosophy, there are some key open problems underpinning a full theoretical understanding of rational agency [403].

2.1.1 Logical Uncertainty

There are different types of uncertainty, with the most common type being the known unknowns of empirical uncertainty [370]. Given significantly less attention than merited by its foundational utility has been logical uncertainty, which addresses the unknown knowns of potentially voluminous entailments of things already known [404]. Just as one can apply deduction about some inductive steps and entailments, one can oftentimes conversely apply induction to expectations about deductions [189, 215, 126, 369]. Logical uncertainty is actually implied, but sorely underanalyzed, in domains such as probability, bayesian reasoning, game theory, and economics, since there are sequences of deductive thinking steps involved [401]. Logical uncertainty bridges rational choice theory and probability theory, combining logic and observation and their respective time-embedded epistemic uncertainties [126, 403]. See section *Logical Induction* on the operationalization and management of logical uncertainty in agents. See section *Open Source Game Theory*, one bridge between intuition pumps and methods for safer self-modification and scalable control that makes use of logical uncertainty. See section *World-Embedded Solomonoff Induction* which would also find progress in logical uncertainty useful. See section *Resource-Aware Reasoning*, one broader context that would be a consumer of this work. See section *Game Theoretic Framing* as more realistic game theoretic models will include logical uncertainty. See section *Grounded Ethical Evolution* as both economic and game theoretic analyses and simulations can be made more realistic when including logical uncertainty.

2.1.2 Theory of Counterfactuals

One key open question in foundational rationality is 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 [405, 113, 403]. See section *Counterfactual Reasoning* regarding drift detection in induction. Also see relevant sections *Logical Counterfactuals*, *Goal Stability*, *Avoiding Reward Hacking*, *Averting Instrumental Incentives*, *Value Alignment*, *Increasing Contextual Awareness*, *Verification*, *Security*, *Oversight*, and *Computational Deference.*

2.1.3 Universal Algorithmic Intelligence

Universal algorithmic intelligence is a theoretical design for an artificial general intelligence that describes a reinforcement learner combining Solomonoff induction and sequential decision theory [214, 213]. Though its original formulation with optimality guarantees has very impactical computational complexity, there are approximations [455] of the algorithm that can be computed, and variants oriented toward more typical AI capabilities [345]. There has even been serious exploration as to making it learn values and be empathetic [344]. It's also used in the seminal measure of algorithmic intelligence, universal intelligence [256]. Because of its theoretical simplicity, this paradigm may be most useful for establishing bounds and theoretical limits of reinforcement learners or artificial agents more generally. See section *Projecting Behavioral Bounds* to which this can contribute.

2.1.4 Theory of Ethics

There are significant open problems in the theoretical and philosophical foundations of both the structure of ethics and its uses.

2.1.4.1 Ethical Motivation

Even if the optimally prescribed ethical theory structure and values are known, and though it may seem obvious to humans, the formal underpinnings and mechanisms of its logical motivation require hardening. An agent that knows what the optimal ethics is would not suffice: it would need to be causally compelled to follow those ethics [402]. See section *Value Specification*.

2.1.4.2 Ontological Value

Within philosophical ontology, there are questions remaining regarding the theoretical underpinnings of the nature of, and treatment of, worth and value. The structural relationships between worth, value, value evaluation, morals, ethical systems, and instantiated ethical systems are insufficiently resolved for the present purposes with long-term AI. Another significant open question is whether there can exist objective values, and if so, how can they be derived. Also see relevant sections *Value Alignment*, *Value Learning*, and *Value Structuring*.

2.1.4.3 Human Preference Aggregation

The viability, structures, and methods by which the preferences and values of multiple people can be aggregated are also open areas of theoretical and immediately practical import [106, 102]. See section *Value Learning* : much of value learning depends on fair and coherent methods for aggregation of preferences and values. Also see relevant sections *Ethical Ensembles* and *Value Structuring*.

2.1.4.4 Infinite Ethics

For any rational agent able to re-derive or weigh ethics, the question of infinite ethics arises, i.e. how to treat the potentially overwhelming significance of existing or potential sentient entities in what is potentially an infinite universe. There are solutions that utilize an agent's uncertainty and solutions that cite causal lightcones, both enabling the accessible to garner most moral attention, but there are still some open questions [60].

2.1.4.5 Normative Uncertainty

Ideally one would want to determine if there are any methods approaching objectivity for enumerating, generating, or ranking value systems. This kind of metaethics explores more first-principles approaches to ethics. Questions about the most proper thing for one to do when they're uncertain about what ought to be done in the first place arise [281, 439]. Figuring out what norms govern uncertainty about normative claims, and how can uncertainty about moral claims can be resolved are both open areas of investigation [281]. Answers to these questions are reasonable baseline proxies to help reason about what sorts of conclusions the operators would come to if they had much more knowledge and much more time to think and develop [403, 58, chap. 13]. These problems are largely still in the realm of philosophy rather than computer science, but interdisciplinary approaches would seem to be be called for in order to explore different angles on this problem. See section *Degree of Value Evolution* as that is also about uncertainty with respect to which values to value, but from the perspective of purposely allowing evolutionary dynamics in them.

2.1.5 Bounded Rationality

There remain foundational questions within the fields of logic, deduction, and philosophy about the theoretical underpinnings of what it means to be rational when there are insufficient resources to satisfy standard rationality [363, 192]. Also see relevant sections *Logical Uncertainty* and *Resource-Aware Reasoning.*

2.2 Consistent Decision Making

To ensure that decisions made are not erratic, unstable, or suboptimal, explicit constructs to allow for stable, consistent, reasonable, and ideally optimal decisions given the knowledge, goals, and faculties available to the agent for it to be robust [371].

2.2.1 Decision Theory

Decision theory, the mathematical study of strategies for optimal decision-making between options involving different risks or expectations of gain or loss, has been long studied [336], but existing methods are not robust, especially with regard to counterfactual reasoning. Mathematical tools such as formal logic, probability, and decision theory have yielded significant insight into the foundations of reasoning and decision making. However, there are still many open problems in the foundations of reasoning and decision [405], and solutions to these problems may make the behavior of very capable systems much more reliable and predictable [371, 46]. Developing a general theory of highly reliable decision-making, even if it is too idealized to be directly implemented, gives us the conceptual tools needed to design and evaluate safe heuristic approaches [403, 405].

2.2.1.1 Logical Counterfactuals

Consideration of hypothetical scenarios in which logical realities and consequences were different than they currently are can and should be evaluated to support causal understanding, understanding of current options, and to enable the generalization of regret terms in learning algorithms. In a counterfactual situation where a given deterministic decision procedure selects a different action from the one it selects in reality, how can one determine the implications of this counterfactual on other algorithms? [403] I.e., is there some satisfactory way to formalize and generalize logical counterfactuals? [403] A method for reasoning about logical counterfactuals in the first place seems a prerequisite to formalizing a more general theory of counterfactuals [405]. While updateless decision theory consistently theoretically outperforms causal decision theory, it will remain an abstraction until such formalizations of counterfactuals are developed [204, 247]. Relatedly an analysis of Lob's theorem in bounded scenarios has implications for cooperative behavior [113]. See section *Counterfactual Reasoning.*

2.2.1.2 Open Source Game Theory

Game theory becomes qualitatively different when players are translucent [190] or transparent to each other. Open source game theory, optimal decision making in a multiagent environment where agents can see each other's source code, is useful for the modeling of one agent controlling another [113, 441]. One key goal of this area is to find ways to foster robust cooperation between agents [39, 403]. It is also particularly useful for scenarios where one agent will be creating another agent that is to be trusted. See section *Controlling Another Algorithm* to which open source game theory has direct import. Also see relevant sections *World-Embedded Solomonoff Induction* and *Logical Uncertainty.*

2.2.2 Safer Self-Modification

Sufficiently advanced agents may modify themselves. Doing so in a stable manner and without leading to detrimental directions is a challenge [403]. Though very difficult to get right, there are some techniques to help make self-modification or generation of successors to be more goal-stable. Early attempts at making this safer leverage extensive introspection and value-based prioritization [317, 420, 421]. But because this self-improvement dynamic can snowball, one would like stronger assurances around safety [403]. See section *Controlling Another Algorithm* as "self" modification tasks are sometimes interested in creating successor agents or alternatively subagents, which should typically be subject to control.

2.2.2.1 Vingean Reflection

If a system will attain superintelligence through self-improvement, then the impact of the system depends entirely upon the correctness of the original agent's reasoning about its self-modifications, and the subsequent ability of its successor to do the same [142, 403]. Generally when one requires extremely high reliability today, implementation-verifiable formal logic based systems are used. Formalized logic to reason about the correctness of self-modifications or prospectively generated superior successors is termed Vingean reflection [142].

2.2.2.1.1 Abstractly Reason About Superior Agents Reasoning about and prediction of what the prospective superior agent would do requires a more viable approach than computing all possible action paths [481]. Agents must abstractly reason about agents which are smarter than themselves, without attempting to replicate the full processes of the superior agent, highly leveraging a rich landscape of counterfactuals [456, 142].

2.2.2.1.2 Reflective Induction Confidence Establishing the confidence in such reflective induction can be aided by using logical induction to establish a framework for determining confidence when formally reasoning about superior agents in the first place [403]. See section *Logical Induction*.

2.2.2.1.3 Löbian Obstacle The Löbian obstacle, essentially that no sufficiently strong formal system can know that everything that it proves to be true is actually true, is a formidable issue when considering recursive verification [403]. An important question to resolve is how can agents gain very high confidence in agents that use similar reasoning systems while avoiding paradoxes of self-reference [142].

2.2.2.2 Optimal Policy Preservation

One technique for stability of goals for nearer-term systems involves creating the conditions under which modifying the reward function of a markov decision process would preserve the optimal policy [309]. See section *Goal Stability*.

2.2.2.3 Safety Technique Awareness

As a system gains skill in software development and algorithms research, explicit modeling and usage of the variety of safety techniques listed in this document can help it to improve its overall robustness and goal stability. See section *Ethical Motivation* as it will need to not only understand these techniques but want to apply them. See section *Metareasoning* as reasoning about the agent's internal processes with cognizance of these techniques may aid long-term robustness.

2.2.3 Goal Stability

The maintenance of stability in the objectives of an advanced agent is challenging [246, 305]. See section *Metareasoning*, types of which enable an agent to choose to e.g. avoid wireheading [246]. Methods for safer self-modification can be used to help maintain goal stability in systems that self-improve. See section *Vingean Reflection* which is key to this. See section *Avoiding Reward Hacking* which is a crucial prerequisite to goal stability. See section *Corrigibility* for which goal stability is a prerequisite. When a static objective is acceptable, e.g. for some for short-to-medium-lived agents, one would want mechanisms to maintain the stability of the overall goals while providing flexibility as to situationally appropriate subobjective priority. See section *Multiobjective Optimization* for this purpose, as multiobjectively optimized ensembles of subobjectives containing specifically formulated goal stability subobjectives might prevent reward function modification. See section *Degree of Value Evolution* which considers why one might want evolution of goals and how that might be managed.

2.2.3.1 Nontransitive Options

Whether individual operators, an aggregation of values of a large number of humans, or potentially the agent itself can have nontransitive preferences, where cycles or loops of comparative preferences occur [133]. While some approaches seek to eliminate such a situation before it arises, more fault-tolerant approaches will attempt to handle such cycles gracefully [316], and this is an open area of research. It is conceivable that solutions to problems like low impact and corrigibility will result in agents that violate one or more Von Neumann-Morgenstern axioms [435], and neuromorphic architectures such as the common deep learning architectures of the day, will almost certainly violate them. The challenge is to do this in a way that's reflectively stable, having the agent not want to rewrite itself into a different agent, and still allows the agent to have a sensible world model. Minimax decision rules are one plausible variant where this can hold [435].

2.3 Projecting Behavioral Bounds

In order to assess risks, it is helpful to understand the theoretical limits on the capabilities being considered [385]. One can understand the theoretical upper and lower limits of a class of agents without needing to know the details of their architecture [486]. For advanced agents, the methods by which developers or operators can determine the theoretical bounds or maximum capabilities of a system include proofs stemming from its architecture and communication of the state of the system. See section *Open Source Game Theory*.

2.3.1 Computational Complexity

It may be possible to characterize or bound the time or resources particular algorithms or capabilities would require, both in their exact and approximate forms [385]. Understanding upper, average, and lower time and resource bounds regarding particular types of functionality would be useful for purposes of projecting risks and weighing tradeoffs [240]. Typically such analyses are done by algorithm designers upfront, but enabling advanced systems to do this analysis well themselves can help tighten the estimates of resources required for specific actions, leading to more robust decisions. See section *Resource-Aware Reasoning* about such requirements.

3 Verification

Verification is a class of techniques that help prove a system satisfies particular properties its designers desire. They provide high-confidence assurances that a system will satisfy some set of formal constraints, helping to answer whether the system was built correctly *given its specification.* [371] When possible, it is advisable for systems in safety-critical situations, such as medical devices or weapons, to be verifiable. Verification of AI systems poses additional challenges on top of those in verifying more traditional software, but It should be possible in many cases to verify the designs of AI systems. If such systems become increasingly powerful and safety-critical, verifiable safety properties will become increasingly valuable and worthwhile.

See section *ML with Contracts* because when whole AI systems and agents are designed by contract, piecewise verification becomes more tractable. See section *Monitoring* because when causal flows are more explicit, verifiable AI systems will also lend themselves to better interpretability and transparency. See section *Decision Theory* as the definition of correct or optimal behavior, among the good levels to verify, is made more explicit by it.

3.1 Formal Software Verification

When people desire extremely high reliability, e.g. for autopilot software, they often use formal logical systems to maximize their certainty of implementation correctness [132, 371]. This is the correct-by-construction approach to software engineering, where a system is developed in tandem with a detailed formal specification and a proof of total correctness given that specification, usually by generating the system from the formal specification [250]. Creating a provably correct implementation, given a specification, is applicable for a range of layers of the software stack [150, 35, 95]. The seL4 kernel, for example, is a complete, general-purpose operating system kernel that has been mathematically checked against a formal specification to give strong guarantees against crashes and unsafe operations [235]. For systems or agents that operate in environments that are at best only partially known by the system designer, it may still be practical to verify that the system acts correctly *given the knowledge that it has* , which avoids the problem of modelling the real environment [127] but puts much stronger onus on the formal specification to be valid. See section *Verified Downstack Software* since having the lower layers of a system being verified often gets much of the benefit with appreciably less overhead.

3.1.1 Verified Component Design Approaches

The assembly of more compound software that combines pre-verified components greatly eases the use of correct-by-construction techniques [371]. If theories of extending verifiable properties from components to entire systems hold, then even very extensive systems can hold certain kinds of safety guarantees, potentially aided by techniques designed explicitly to handle the higher-level semantic, behavioral, and distributional properties [373] of AI and ML systems. Modular systems of formal mathematics [282, 361] can serve both as a model for, and as a resource in creating, such modular verified frameworks and components. See section *ML with Contracts* since those techniques can help significantly with architectural componentization decisions. See section *Careful Engineering*, since applying modular design, verification, and other careful process to at least the base of an AI system can mitigate many danger points.

3.1.2 Adaptive Control Theory

Systems that interact with the external world, such as cybersecurity systems and trading agents, necessarily have parameters that vary or are initially unknown. The field of adaptive control theory addresses design of such systems and serves as a step in the direction of the AI systems being addressed in this landscape guide [29]. Some work has been done on verification of such systems [111], and may help inform verification of more sophisticated learning systems [371].

3.1.3 Verification of Cyberphysical Systems

When the external world that agents interact with is the physical world, further complications come into play. It is often difficult to actually apply formal verification techniques to physical systems, especially systems that have not been designed specifically to support verification. This has motivated research seeking a general theory linking functional specifications to physical states of the world [343], usually confined to the area surrounding the system. This type of theory would allow use of formal tools to anticipate and control behaviors of systems that approximate rational agents, alternate designs such as satisficing agents, and systems that cannot be easily described in the standard agent formalism (powerful prediction systems, theorem provers, limited-purpose science or engineering systems, and so on) [371]. Such a theory may also allow rigorous demonstrations that a system is prevented or constrained from performing certain kinds of reasoning or taking certain kinds of actions [371]. The application of reachability analysis and hybrid discrete/continuous dynamics to the formal verification of computational physical entities operating in physical spaces seem promising [343, 7, 465, 222, 277, 296, 297, 8], as does extending the introspection and proof capabilities of differential dynamic logic [154].

3.1.4 Making Verification More User Friendly

Creating verified or correct-by-construction software has a much higher overhead as compared with unverified software [330]. Creating methodologies, algorithms, tools, and cultures around creating verified software with the aim of making it easier for software developers to decide to use these methods would be quite worthwhile. While verification based on theorem provers can scale to larger systems than ones based on model checkers can, and the former can also reason inductively where the latter is unable to [330], theorem provers pose usability issues for systems of realistic size. The proofs created by such verification systems for software of practical size are unfortunately quite unweildy [330], calling for easier ways to explore, summarize, check, and query them. There are also a number of relevant open problems with proof verifiers [470] that can impact their use on AI.

3.2 Automated Vulnerability Finding

Most systems today are not developed with correct-by-construction methodologies but rather are programmed directly. While techniques for attempting to verify such systems exist, there are many aspects and constructs they cannot cope with, [116] often due to their dynamicity, but even some formal proofs may be unverifiable [470]. Given such less-formally specified systems, automated bug and vulnerability finding by abstract static analysis [116] and static application security testing [65] in conjunction with dynamic testing, is a next best technique for reducing implementation errors. See section *Implementation Testing* for more on dynamic testing. See section *Red Team Analysis*, which addresses vulnerability finding with blackboxes rather than whiteboxes.

3.3 Verification of Intelligent Systems

Additional challenges with the verification of artificial intelligence and machine learning software come from the fact that behavior is more determined by data as compared to concentional software. To verify such systems, one needs not only provably correct implementations of the system's algorithms (given specifications of those algorithms), but also of the models they learn [386]. : introspective environment modeling, end-to-end specifications and specification mining, developing abstractions for and explanations from ML components, creating new randomized formal methods for systematically generating realistic data, and developing verification engines more oriented around run time, quantitative operations, and learning [386].

3.3.1 Verification of Whole AI Systems

The verification of whole AI systems has been underexplored, but a number of challenges are clear. A number of principles to address these challenges have been proposed [386] : introspective environment modeling, end-to-end specifications and specification mining, developing abstractions for and explanations from ML components, creating new randomized formal methods for systematically generating realistic data, and developing verification engines more oriented around run time, quantitative operations, and learning [386]. A number of challenges to verifying AI systems in the general case remain [470]. See section *ML with Contracts* as software interface contracts make segmentations within the analyses needed for formal verification much easier. Following a componentized architecture, in which guarantees about individual components can be combined according to their connections to yield properties of the overall system seems promising [386]. Review and remediation of systemic issues such as questionable data dependencies, unpleasantly unexpected data flow dynamics, inappropriate or suboptimal abstractions, and mixed technology stacks [384] will help to simplify AI systems to help make them more likely to be meaningfully verifiable.

3.3.2 Verification of Machine Learning Components

Each type of machine learning algorithm or paradigm may need to be considered individually with respect to what aspects need better formalization, quantization, introspection, or proofs [371]. There has been some work on requirements for verifying neural networks [349, 383]. There has also been exploration of directions for more nuanced and efficient summarization of the high-dimensional distributions many learning algorithms produce [4, 210, 232], whereby quantization of the distributions might enable simple deductive proofs and satisfiability analyses. State coarse-graining or abstraction for reinforcement learners conforming to specified partial programs provide a path for formalized treatment of learned policies [18]. See section *Defined Impact Regularizer* as there are conceptual parallels between partial programs and distance from a baseline policy. In addition to accounting for particular learned models, one must also account for future changes to the learner as new data arrives, another underexplored challenge [386].

3.4 Verification of Recursive Self-Improvement

A related verification research topic that is a distinctive long-term concern is the verifiability of systems that modify, extend, or improve themselves, successors, or descendant programs, quite possibly many times in succession [170, 456, 472]. Attempting to straightforwardly apply formal verification tools to this more general setting presents new difficulties [371]. With many potential approaches, a formal system that is sufficiently powerful cannot use formal methods in the obvious way to gain assurance about the accuracy of functionally similar formal systems, due to Gödel's incompleteness theorems [141, 460]. There has, however, been recent progress in modeling theorem prover verification systems in themselves, specifically illustrated with model polymorphism, cleverly circumventing the Gödellian obstacle [140]. See section *Safer Self-Modification* as it relates safer, if not provably safe, self-improvement, potentially doing so as a style of learning. See section *Vingean Reflection*, which addresses theoretical challenges to reflective proofs.

3.5 Implementation Testing

Typical dynamic testing of software implementations include unit testing, integration testing, functional testing, system testing, stress testing, performance testing, and regression testing. These same kinds of tests can and should be applied to AI systems and agents [384]. While the levels of indirection concomitant to AI mean their domains and ranges are much too large to test exhaustively or even representatively, there remains value in this mode of quality assurance. Though not ostensibly representative of the distributions of what the agent can encounter, unit tests, for instance, can catch some representative failure modes that can be as elaborate and appropriate to the agent's complexities as can be imagined [96]. In contrast to the central role unit testing takes with most software development processes, it cannot be relied on to provide assurance for AI agents, and especially not advanced agents that support multiple layers of indirection. When verifying physical systems, or more generally any system properly subject to an open world model, particular cognizance needs to be paid to validating that the specification is complete.

4 Validation

Given even verified software, environmental assumptions can easily not hold in the real world, or the requirements that led to the specification may be faulty or lacking. These sort of specification errors are quite usual in the world of software verification, where it is often observed that writing correct specifications can be more difficult than writing correct code [371]. Validation is the process of ensuring that a system that meets its formal requirements does not have undesirable behaviors or consequences. It is asking the question "Did I build (or ask for) the right system?".

Ensuring that the formal requirements, the specification, considers all relevant dynamics, and will be beneficial and desirable, does not actually fit into current formally provable paradigms. In order to build systems that robustly behave well, one currently needs to decide what *good* behavior means in each application domain. This ethical question of good is tied closely to questions of what technologies and engineering techniques are available, how reliable they are, and what trade-offs can be made — all areas where computer science, software engineering, machine learning, and broader AI expertise are valuable. In practical application, a significant consideration is the computational expense of different behavioral standards or ethical theories, i.e. that if a standard cannot be evaluated sufficiently expeditiously to guide behavior in safety-critical situations, cheaper approximations should be used [458]. It is therefore likely best for different complexities of ethical or moral reasoning to be used for different timescales.

Designing simplified rules, such as those meant to govern a self-driving car's strategic decisions in critical situations, will likely require expertise from both ethicists and computer scientists [371]. This is relatively straightforward when specific safety limitations, behaviors, and ethical constraints are known upfront for AI systems that are largely specified upfront [48, 451], but more powerful and operationally-flexible systems require much more sophistication. Computational models of ethical reasoning may shed light on questions of computational expense and the viability of reliable ethical reasoning methods [26, 424]. Being able to assure known bounded behaviors from methods and systems that learn will also be important for both medium-term and long-term safety. Validation encompasses ensuring an agent understands its environment, decisions, and actions, and that it acts robustly-in-accordance with its operators' wishes.

In the long term, AI systems might become much more powerful, general-purpose, and autonomous, a regime where failures of validity would carry significantly higher costs than with today's systems. To maximize the long-term value of validity research, machine learning researchers might focus on anticipating, preventing, detecting, and mitigating the types of otherwise unexpected generalization that would be most problematic for very general and capable AI systems. If some concepts could be learned reliably, it might become possible to use those to define tasks or constraints that minimize the chances of unintended consequences even when agents become very general and capable. This topic has been underexplored, so both theoretical and experimental research on it may be useful [371].

4.1 Averting Instrumental Incentives

It has been argued that intelligent systems that aim toward some objective have emergent instrumental incentives, pressures to create subplans, to give them more time, power, and resources to fulfill their objectives [321]. These instrumental pressures, and the flexibility to act on those pressures, are relatively uncommon in today's rather narrow systems, but would seem much more expected and pronounced as AIs become more general [58], and understand a wider array of contexts, modalities, and fields. Progress on formalizing this class of problems has only been recent [47]. How can one design and train systems such that they robustly lack default incentives to manipulate and deceive the operators, compete for scarce resources, prevent their maintenance [321], and generally avert negative instrumental behaviors? [437, 58] Averting these implicit pressures in a design is much more difficult than it may appear initially, and there are numerous subtleties to consider [22, 21, 47]. Sophisticated techniques to avoid such incentives can be needed not only for agents but also oracle-style question answering AIs [19]. See section *Operator Value Learning*, where systems can avert instrumental incentives by maintaining uncertainty over the truly desired goal. See section *Switchable Objective Functions* whereby faulty interpretations of primary objective can be safely corrected without the default instrumental pressure to prevent such a change. Also see relevant sections *Multiobjective Optimization* and *Computational Humility.*

4.1.1 Error-Tolerant Agent Design

A hallmark of robustness is fault-tolerance. In the case where value alignment or design-time effects are flawed, one might not see those issues in an advanced agent until it has been in production a while. Highly capable agents can also be dangerous if they are specified incorrectly [478]. General advanced agents may very well be aware of attempted changes to it, and by default can have emergent or implicit pressures to prevent such changes [403]. Making intelligent systems such that they are amenable to correction, even if they have the ability to prevent or avoid correction, is therefore necessary [403, 437, 326]. Also see relevant sections *Corrigibility*, where preliminary progress on particular types of instrumental incentive aversion have been made and *Utility Indifference*, where control of the objective function is facilitated by the agent's explicit indifference about its utility function.

4.1.2 Domesticity

The optimizers of today do not annex or use excessive resources not because doing so wouldn't lead to more optimal solutions to the function they're optimizing, but because they lack the capability to annex resources. This will change as AI becomes more general in its capabilities. It has been argued that explicitly and safely incentivizing such an intelligent agent to be low-impact on its environment will therefore be necessary [23, 403].

4.1.2.1 Impact Measures

Methods to quantify the amount an agent does, or can do, to change the world become necessary in such regimes. Using such measures, and other techniques, one can start exploring what sorts of mechanisms, including regularizers, might incentivize a system to pursue its goals with minimal side effects [11].

4.1.2.1.1 Impact Regularizers Regularizers, methods for structurally shaping learning or penalizing undesirable learning, can be made to penalize an agent for recognizably changing the world [437].

4.1.2.1.1.1 Defined Impact Regularizer If one has enough of an object-level understanding of what the agent will encounter in advance, one can elect particular impact measures to use in the regularizers that temper impact. One can choose to penalize changing the environment overall [11, 437] or one can introduce a penalty for changes relative to some baseline environmental state or a baseline policy determined by some explicit method [25]. One can start from safe policy and try to improve the policy it from there, in manners similar to reachability analysis [279, 295] or to robust policy improvement [217, 11, 315].

4.1.2.1.1.2 Learned Impact Regularizer Instead of using predefined measures, an agent can learn transferrable side effects across multiple tasks in similar environments, [11] similar in mechanism to transferring just learned dynamics [438]. This would help it learn to characterize and quantify expected, relevant, and unexpected environmental changes.

4.1.2.1.2 Follow-on Analysis Causal analysis of downstream effects can be used to either prune actions that would cause deleterious effects, if valence can be ascertained, or prune excess effects at all, if it cannot [437, 336]. Also see relevant sections *Lengthening Horizons*, *Values-Based Side Effect Evaluation*, and *Action and Outcome Evaluations.*

4.1.2.1.3 Avoiding Negative Side Effects If possible, one should aim to ensure that the agent won't disturb the environment in specifically *negative* ways while pursuing its goals, as opposed to in any way at all, and we'd ideally like to do so without specifying what not to disturb.

4.1.2.1.3.1 Penalize Influence One way to prevent future abuse of power is to minimize the amount of power needed to accomplish the prescribed task, i.e. penalizing empowerment or some other metric of being in a position to cause side effects [376, 11]. See section *Defined Impact Regularizer* to contrast penalizing changing things versus penalizing the ability to change things. As this is essentially the opposite of the type of objective function often used for intrinsic motivation [298] further study is needed on viable formulations of the idea. It has however also been noted that reinforcement learning techniques which punish the agent for attempting to accumulate control would actually incentivize the agent to deceive and appease its creators or operators until it found a way to gain a decisive advantage [403, 58, chap. 8].

4.1.2.1.3.2 Multi-agent Approaches Cooperative multiagent techniques may also disinventivize negative side effects. Transparency of, and cooperation toward, a shared reward function by multiple agents will continually solicit input from multiple stakeholders, and in so doing reduce risk of undesired outcomes per their varied models [11]. For example, biasing toward goal transparency [180] via autoencoding reward functions, or alternatively cooperative inverse reinforcement learning, across multiple agents, can help ensure they're happy with resultant environmental changes [186]. See section *Cooperative Inverse Reinforcement Learner*, a cooperative, though assymetric, multiagent algorithm.

4.1.2.1.3.3 Reward Uncertainty Aware that random changes to the environment are more likely to be bad than good, one can maintain and evaluate a distribution of reward functions, optimized to minimize impact on the environment [11]. These different sub-reward-functions can each be sensitive to different kinds of undue impacts on the environment, or even just by virtue their variety cause noticable negative side effects to be less common. See section *Operator Value Learning* which maintains a distribution of different operator-intended objective functions. See section *Penalize Influence* as implementation paths are similar. See section *Resource Value Uncertainty* for a resource-differentiated analogue. See section *Multiobjective Optimization* as that provides additional degrees of freedom to subobjectives and should also achieve the target effect.

4.1.2.1.4 Values-Based Side Effect Evaluation More explicit approaches to avoiding negative side effects, combining common sense reasoning and evaluation of scenarios with respect to values, offer an integrative approach with other aspects of safety and alignment [420, 180]. Such techniques can evaluate prospective side effects based on core, compound, and instrumental values in a consequentialist manner. Also see relevant sections *Value Alignment*, *Follow-on Analysis*, and *Action and Outcome Evaluations*.

4.1.2.2 Mild Optimization

AI has typically been about optimizing for a given function as much as possible, but with power and intelligence, functions can be optimized much further than people want or need when not paying heed to side effects [437]. One would therefore like to be able to achieve a semi-optimization of a target function with relaxed optimality requirements, such as by near-optimization or approximate optimization.

4.1.2.2.1 Optimization Measures Varied ways of measuring optimization and penalizing excessive optimization are emerging. One can regularize against excessive optimization power [437], learning corresponding value functions for policies in order to learn less-extreme policies that are more likely to generalize well [146], or just generally provide less optimization power [350]. Another option is a regularizer that penalizes more intelligence than is necessary for solving a problem sufficiently, biasing toward the speed of solution and resource conservation [437]. Penalizing or thresholding the amount of resources or time that can be used for optimization, or stopping when reaching some probabilitsic theshold of the optimum, are options [236, 376], but each has potential issues, and optimization measures themselves are an open problem [256].

4.1.2.2.2 Quantilization Rather than seeking the most extreme solution to a problem, i.e. maximizing a function such as expected reward, one can instead seek to satisfice expected reward [396, 437]. This may be approached by choosing actions from a top percentage of possible actions per a sort by expected reward or probability of success though that alone wouldn't necessarily prevent satisfaction from overoptimization [436]. The technique of quantilization selects a strategy randomly within some top percentile of strategies [436], which probabilistically mitigates such risk. See section *Logical Counterfactuals* as that can assist with enumeration of options.

4.1.2.2.3 Multiobjective Optimization The simultaneous optimization of two or more subobjectives enables soft consideration and formalized tradeoff of resource usage and other types of impacts, including as measured with respect to arbitrary specified objectives [124, 264, 254, 284]. See section *Nontransitive Options* as there are open issues with integrating multiobjective optimization into safe advanced AIs that conform to the Von Neumann Morgenstern axioms due to the possibility of intransitivities arising in iterated pareto frontiers [99]. See section *Multiobjective Reinforcement Learning* for treatment in an RL context. See section *Multiple Rewards* as that considers an analogue where subobjective variants aim to represent the same conceptual goal. See section *Ethical Ensembles* as those can be implemented using multiobjective optimization, and further may gracefully balance multiple value systems and multiple operational objectives simultaneously.

4.1.2.3 Safe Exploration

In addition to various types of resource exploitation being potentially unsafe, exploratory moves can also be unsafe if they bring the agent into an unsafe situation. In order to help ensure that the agent doesn't make exploratory moves, in order to learn, that have very bad repercussions, a variety of safe exploration techniques have been discussed [11, 160, 338]. See section *Trusted Policy Oversight* as that can enable safe exploration given blessed core control [11]. See section *Distance from User Demonstration* as the state space surrounding those is typically a safe space to explore. See section *Scalable Oversight* as the oversight can include clarifications of which areas are safe to explore.

4.1.2.3.1 Risk-Sensitive Performance Criteria One can change optimization criteria from just total expected reward to other objectives that prevent or minimize downsides of note [160, 432, 11, 442]. It's also possible to estimate uncertainty within value functions, which could be incorporated into risk-sensitive reinforcement learning, [329, 156] which may also model "intrinsic fear" with probabilities on how close in steps different states are to catastrophe [272].

4.1.2.3.2 Simulated Exploration One way to explore more is to explore in simulations as much as possible. Agents can then safely incrementally update the policies from imperfect simulation-based trajectories with sufficiently accurate off-policy trajectories using semi-on-policy evaluation [11]. Such techniques can also apply in senses broader than strictly exploration [85].

4.1.2.3.3 Bounded Exploration Another way to explore safely is to model or otherwise determine the state space regions in which mistakes are relatively inconsequential or recoverable, and exploring only within those areas [299, 11]. Recent techniques that do this iteratively for safe exploration using markov decision processes [449] and gaussian processes [381] can explore unknown environments without getting into irreversible situations. See section *Trusted Policy Oversight* as bounded exploration may be a useful component of that.

4.1.2.3.4 Multiobjective Reinforcement Learning Multiobjective reinforcement learning is a generalization of reinforcement learning extended to multiple simultaneous feedback signals [274], and this can include as one of its subobjectives minimizing riskiness.

4.1.2.3.5 Human Oversight One can have agent determine which exploratory actions are risky, or are ambiguously risky, and query a human about them, making known-safe decisions until the human responds [11]. See section *Oversight* for scalable methods designed specifically for control.

4.1.2.3.6 Buried Honeypot Avoidance One technique for learning safe exploration involves allowing monitored bugs and vulnerabilities to exist (in an evaluation environment) that can be exploited to achieve a higher reward (or other optimization measure), and selecting algorithms or agents that generalize to not exploiting such unintended and unconventional avenues [437], at least for some classes of them. See section *Tripwires*, which are similarly structured but are meant to be used later in the agent life cycle.

4.1.2.4 Computational Humility

Algorithmic techniques enabling an agent to humble itself [414] and moderate its views of its own importance can be termed computational humility. See section *Perceptual Distance Agnostic World Models*, which can aid humble knowledge representation.

4.1.2.4.1 Generalized Fallibility Awareness One central type of deference for advanced agents is to know that it may be that none of its hypotheses model the world well enough, and additionally biasing to consider its operators are less flawed than it is [437], so letting them shut it down or modify it if it seems like that's what they want to do. See section *Computational Deference*, which includes the very related control-centered aspects of corrigibility. An agent can try to model all of the ways it's flawed, but that may require that those models not be discounted by their own mechanism [437], an open issue.

4.1.2.4.2 Resource Value Uncertainty Maintaining dynamic bayesian uncertainty regarding the importance or value of unexplored and underexplored or uncharacterized entities, relationships, and dynamics can significantly effect humble treatment of others [414, 285]. Reasonable multilevel aggregation would be needed to deal with constructs of varying temporal, physical, and complexity scales [398, 284]. Also see relevant sections *Multilevel World-Models* and *Action and Outcome Evaluations*.

4.1.2.4.3 Temporal Discount Rates Resource acquisition strategies can be tempered by the effects of applying a large temporal discount rate on the expected reward [371].

4.1.3 Averting Paranoia

Another instrumental pressure is to constantly scan or prepare for possible ill-intent or possible information leakage. While security precautions and vigilence are important, an arbitrarily high amount of resources can be spent on threat detection and avoidance, while that should typically be tempered to be sufficient to prevent and deal with realistic threats [98]. See section *Computational Deference*, on techniques for having an agent trust its operators. See section *Mild Optimization* for more on optimizing objectives and subobjectives to an appropriate amount but no more. See section *Handling Improper External Behavior* against which this needs to be balanced.

4.2 Avoiding Reward Hacking

A very intelligent agent built to optimize just its observations rather than some function in the actual environment would likely not align with human interests [58, chap. 12]. It may very well cause unanticipated and potentially harmful behavior by gaming its reward function, deluding itself whether purposefully or inadvertantly [444, 54]. Determining how one can reliably prevent an agent from gaming its reward function like this is an open area of research, thus far with a number of promising ideas for mitigation [128, 327]. See section *Tripwires* which may be used to help catch these behaviors. See section *Ontology Update Thresholds* as the dynamic degradatory repurposing of properties as in Goodhart's Law surfaces in both [58].

4.2.1 Pursuing Environmental Goals

A key question in AI safety is how one might create systems that robustly pursue goals defined not in terms of their sensory data of their environment, but in terms of the state of the actual environment [437]. Behaviors of self-delusion, addiction, and paying attention to unimportant fictions where it seems like reward can be generated are all termed wireheading [354]. It has been a noted issue in counterfactual learning [62, 427], in contextual bandits [5], and will very likely become more common as environments grow in complexity [11].

4.2.1.1 Environmental Goals

Environmental goals refer to the structuring of a reward system such that an agent that would fool its sensors into registering otherwise-high-reward data would not actually receive a high reward [128, 201, 437, 138, 129].

4.2.1.2 Predicting Future Observations

One way to avoid self-delusion is to abductively generate, maintain, and test multiple hypotheses of the causes of received sensory input in order to distinguish illusory states from environmentally true and valid states, and update priors regarding past sensory data [437].

4.2.1.3 Model Lookahead

Model lookahead can also be used to base reward on anticipated future states rather than the current state, to penalize and discourage the agent from modifying its reward function [11, 139, 201].

4.2.1.4 History Analysis

An agent's historical trajectory can be analyzed for incongruences, discontinuities, and hints of efforts to self-delude, and if delusion is detected this can trigger an appropriate backtracking [437].

4.2.1.5 Detecting Channel Switching

The entry of an agent into wireheading may be indicated by percept disjointedness. Delusional hallucination might therefore be avoided using "slow features" to detect incongruences or discontinuities that are indicative of a percept-channel switching event [467]. This involves the application of anomaly detection to normally slow-changing features to detect such segmentation [437]. Note that this issue is very different from either the hallucinations of "inceptionism" [301] used for transparency into deep networks or the sorts of imaginings that generative networks [485] do to produce exemplars.

4.2.2 Counterexample Resistance

One might design a system to build resistance [11] to natural and unnatural counterexamples, confounding patterns, and adversarial tricks by leveraging adversarial training [171, 56, 147]. This discourages the agent from purturbing its data to achieve alternate interpretations. See section *Adversarial ML* as this can also be framed as a security issue when external parties are providing the adversarial testing or production data.

4.2.3 Adversarial Reward Functions

In order to enable a more active, critical, and responsive reward function, one can have the reward function itself be an agent that tries to find flawed value determinations, similar to how generative adversarial networks work [11, 172].

4.2.4 Generalizing Avoiding Reward Hacking

Independent of specific techniques, formalization and study of reward hacking by analyzing the distance and shape of jumps in the feature space would help to generalize its detection and avoidance [437].

4.2.5 Adversarial Blinding

Adversarial techniques can be used to blind a model to certain variables, useful for masking how an agent's reward is generated, and enabling cross-validation for agents [11, 159].

4.2.6 Variable Indifference

By maintaining the independence of variables actually valued, the system can shape the direction of the optimization pressures to areas one cares less about maintaining to moderate values [11].

4.2.7 Careful Engineering

To reduce the risk of reward function gaming, one can utilize extensive verification, testing, and security to create a safer base or core agent, on top of which more specific agents can be configured. See section *Containment* as one can be further careful about isolating the agent from its reward signal using that [11].

4.2.8 Reward Capping

In order to prevent extremely low-probability high-reward choices, carefully normalizing rewards or utilities at a level appropriate to the agent [11] is in order. Concomitantly, model uncertainty should dominate most expert calculations that involve small probabilities [325].

4.2.9 Reward Pretraining

Another way to discourage gaming of the reward function is to train the reward function in a supervised manner offline, ahead of online use [11, 149]. As a known function, it can be analyzed analytically, statically, or in unit tests, and it will by definition be robust against online reward function corruption.

4.2.10 Multiple Rewards

Another reward function robustness method is to use an aggregate of different variants or proxies of the same basic objective [11, 124]. See section *Knowledge Representation Ensembles* as that provides variants of the objective function for aggregation via varied world models. See section *Multiobjective Optimization* as that uses similar techniques but can also include independent subobjectives. See section *Multiobjective Reinforcement Learning* for treatment in an RL context.

4.3 Value Alignment

Researchers expect that highly intelligent agents would be able to pursue arbitrary end-goals that are independent of moral values, since intelligence and values are hypothesized to be orthogonal [59]. Values are ethical norms, constraints, and moral weights one places on properties and relationships in the world. Ensuring that a system conceptualizes and uses its intelligence in a **beneficial** manner requires more than accuracy [402]. It has been argued that a single fixed pithy solution to machine ethics may be rather hole-ridden, brittle, and deficient in fairness [67]. Even Asimov's deontological scenarios illustrate such loopholes, thus the need much more comprehensive paradigms [462]. As systems become more capable, more epistemically-difficult methods of value loading could become viable, suggesting that research on such methods could be useful [58]. A portfolio approach is warranted to support a variety of methods for specifying goals indirectly and semi-indirectly. See section *Informed Oversight* as further research in that control problem may change the field's understanding of the nature of value loading.

4.3.1 Ethics Mechanisms

There are multiple methods to represent such values and to imbue them into artificial agents. Different orders of complexity of AI systems will have different capacities and mechanisms to represent values [416, 284]. Highly capable and general AIs will require more sophisticated treatments of values.

4.3.1.1 Grounded Ethical Evolution

An approach to growing ethical systems is to use mechanism design, setting up an environment with multiple agents, and incentives and punishments, to direct norms and preferences of the agents toward cooperation [68, 107, 359]. Unfortunately, in the limit, such evolution is by no means guaranteed to reach a desirable or context-insensitive state [188]. See section *Game Theoretic Framing* regarding the ability of agents to model each other aid at the arrival to game theoretic and similar equilibria. See section *Open Source Game Theory* regarding similarly framed dynamics but with agents' source code available to each other. See section *Correlation of Dynamics* as the correlations between an agent and its environment, as well as with other agents, largely shape the ethics that evolve. See section *Logical Uncertainty* as both economic and game theoretic analyses and simulations can be made more realistic when including logical uncertainty. See section *Norm Denial of Service* as an example issue that can arise in such systems when realistic timing and thus logical uncertainty are ill accounted for.

4.3.1.1.1 Moral Trade Within such systems, one dynamic of note is that moral agents that disagree about morals will tend to trade to increase their moral impact [324]. Another recent concept of note is ethical fusion, in which aggregation in collective and collaborative decision making occurs not via linear combination but effectively through negotiating synergies [180, 359]. Evolutionary development in general, however, can lead to unexpected results that are undesirably specific to incidental environmental factors [54].

4.3.1.2 Value Specification

Broadly, value specification is the direct or somewhat indirect specifying of values a system should hold and act on. The intentions of operators are, however, fuzzy, not well-specified, and sometimes containing contradictions [480]. Some prominent AI researchers expect moral philosophy to become an increasingly important commercial industry [366]. With advanced agents, it's not sufficient to develop a system intelligent enough to figure out the intended goals, though; the system must also somehow be deliberately constructed to pursue them [58, chap. 8] [403]. Also see relevant sections *Multilevel World-Models*, *Uncertainty Identification and Management*, *Operator Modeling*, and *Normative Uncertainty.*

4.3.1.2.1 Classical Computational Ethics Computational models of ethical reasoning may shed light on questions of computational expense and the viability of reliable ethical reasoning methods [371, 458, 14, 424, 180, 359, 360, 26]. In scenarios where an agent operates in a human-level world, where humans can relate to states and actions, whether physical or virtual, pluggable ethical profiles may be quite appropriate [366]. Machine learning of ethical features and determinations from human-labelled data may be a plausible method of bootstrapping moral features that are relevant for human-like experiences [105, 15].

4.3.1.2.2 Value Structuring How can one structurally, algorithmically ensure that the core values of advanced learning agents meet broad beneficence constraints? Arguments have been made that they should be heavily informed by analyzing how humans structure their value systems [178, 380, 220]. See section *Psychological Analogues*, as these begin to point toward parallels between human psychology and analogues in machine learning.

4.3.1.2.2.1 Values Geometry The way that value-related concepts, relationships, and skills are represented may heavily influence the dynamics of their use. What the basic data structures, or even theoretical ontological structures, of ethical and moral values are is not yet a settled question though. Some schemes model values as worth attached to ontological properties and relationships, providing a range for preferences [180, 359, 284]. There have been computational explorations of how human values should be conceived of and modeled [412, 360]. There may also be cartesian and monte carlo tree search graph search algorithm constructs of values which might be well-suited to efficient ethical ensembles [57, 284]. See section *Concept Geometry* as values consist at least in part of particular kinds of concepts.

4.3.1.2.2.2 Action and Outcome Evaluations Combining retrospective and prospective perspectives would seem a good desideratum for ethical contemplation. One such joint representation and evaluation mechanism for developing and using values is to assign value to choices by a combination of intrinsic value, based on past experience, and from an internally represented causal model of the world [114]. There can also be combinations of hard and soft constraints, conflict resolution, and consistency mechanisms over multiple layers of values, ethics, morals, and preferences [180, 359]. See section *Follow-on Analysis* which analyzes and prunes deleterious downstream actions. See section *Lengthening Horizons* which seeks ways of increasing outcome prediction horizons.

4.3.1.2.2.3 Game Theoretic Framing Putting common notions of morality into a game theoretic framework in order to support propagation of prosocial values and equilibria is a promising, if early, framing [260, 105]. Evolving ethical systems, however, does not yet provide many guarantees as to success or applicability [188], and so research into the differential robustness of such solutions when the mechanism is modified may be beneficial. See section *Logical Uncertainty* as more realistic game theoretic models will account for logical uncertainty.

4.3.1.2.2.4 Unified Ethics Spaces Having a harmonized space in which different major or important ethical systems can be represented, e.g. vector or tensor spaces, or graph traversal spaces, might allow for efficient set theoretic operations across them, facilitating ensemble and multiobjective approaches, but this is rather underexplored as yet [180, 284].

4.3.1.2.3 Capability Amplification Capability amplification, starting with an aligned policy and using it to produce a more effectual policy that is still aligned, offers an interesting paradigm for cyclic active learning [84].

4.3.1.3 Value Learning

It has been argued that it's quite plausible that researchers and developers will want to make agents that act autonomously and powerfully across many domains [371]. Specifying one's preferences somewhat explicitly in broad or general domains in the style of near-future narrow-domain machine ethics may not be practical, making aligning the values of powerful AI systems with one's own values and preferences difficult [402, 403]. Concretely writing out the full intentions of the operators in a machine-readable format is implausible if not impossible, even for simple tasks [403]. An intelligent agent must be designed to learn and act according to the preferences of its operators, likely with multiple layers of indirection [437].

4.3.1.3.1 Operator Value Learning To learn operators' values effectively, it may be necessary to maintain and refine a distribution of possibilities of the actual meaning of those goals or values specified or hinted at [437]. Such a system that believes that the operators (and only the operators) possess knowledge of the "right" objective function might be very careful in how it deals with the operators, and this caution could therefore avert potentially harmful default incentives [185, 402, 144]. Such systems may avert instrumental incentives by virtue of the system's uncertainty about which goal it is supposed to optimize. A major part of this process is the iterative extraction of one or more humans' volition, of which there are a few methods [91, 402]. See section *Value Elicitation*, as that addresses extraction methods more directly. Before particular architectures are selected, considering broad paradigms like reward engineering [88] can help inform an approach. Very advanced AIs should in theory be able to take the preferences, values, and volition from one or many humans and extrapolate forward to the additional entailments and moral progress that additional knowledge and additional time to think would produce [479, 434]. See section *Operator Modeling.*

4.3.1.3.2 Value Elicitation Value elicitation involves finding sources of values and sufficiently interrogating them to extract the values.

4.3.1.3.2.1 Value Sourcing Good people are quite varied in their morality and values [187]. Societal values also change over time. In order to attempt to start with reasonable recall over values, the number of types of artifacts, modalities, stakeholders, sapient species, and organizations providing value laden information might be maximized. Techniques for analyzing consistencies and synergies can then be used to improve precision [359]. 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 will be a practical necessity regardless of which philosophical view is the correct one. The methodology used for such sourcing requires fairness and objectivity [179]. Having a set of human operators with total control over a superintelligent system could give rise to a gigantic new moral hazard, however, by putting historically unprecedented power into the hands of a small few [58, chap. 6].

4.3.1.3.2.2 Value Interrogation The set of methods for querying or detecting values from sources of values is called value interrogation. Depending on the way one wishes learned values to be structured and how directly those should be learned, mechanisms for querying authorized sources of values can appropriately vary considerably, ranging from asking, to watching, to scanning [63, 402, 434, 344]. See section *Value Specification* in which values are queried and provided with less indirection. See section *Value Learning* 's other children, as they include methods for collaborating with or scanning value sources. See section *Robust Human Imitation* in which values are learned by analyzing operator worldlines. Techniques for maximal extraction of alignment-oriented information will also include robust counterfactual elicitation. See section *Theory of Counterfactuals* as hypotheticals and the ranges of valid options are important in characterization of values. The field of preference elicitation [76] likely has much to offer that of the rather related value elicitation.

4.3.1.3.2.3 Value Factoring When values are identified, whether implicitly or explicitly, these may be either instrumental values, narrow short-term values, or terminal fundamental values [80, 344]. Correctly learning terminal values indirectly, ideal theoretically for superintelligence, may require unmanageable amounts of resources [80]. If one applies active learning to disambiguating common roots or causes among different narrow values, in essence prompting humans to perform goal factoring, [75] the system and operators in tandem may be able to align and stitch together the network of values iteratively deeper. With an active inverse reinforcement learning setting driving the elicitation, [24] correspondences and disconnects between observed behavior and stated preferences, goals, or values may be able to be accounted for. See section *Values Geometry* for discussion of how compound and terminal values may be structured.

4.3.1.3.3 Collaborative Values Learning Learning of instrumental or potentially terminal values can be done in a collaborative manner, whereby multiple agents interrogate each other, perform joint trial-and-error episodes, and preemptively disambiguate concepts or events for each other in order to uncover the reward function of a target subset of the agents [185]. See section *Cooperative Inverse Reinforcement Learner* as cIRL is a leading algorithm for this purpose.

4.3.1.4 Ethical Ensembles

As human ethics often contains competing or contradicting tenets, and as there are multiple different structures of semiformal moral philosophy, an ensemble of ethical systems, evaluating multiple different blessed ethical systems in tandem and using some aggregate to inform decisions, may be called for [281, 415, 284, 180, 305].

4.3.1.5 Structured and Privileged Ethical Biases

To strike a balance between specified and learned values, one can structure a mechanism of ethical constraints of variable softness. Core values specified upfront would be rather firm, and instrumental or derivative values, preferences, and actions would be built and learned atop them, and evaluated against them, [433] with incremental layers featuring incrementally softer and more malleable constraints [420, 284, 459]. See section *Privileged Biases* as the agent can not only ignore or reject things that go against its core values, but can focus on those as potential threats or tracked as sources of issues.

4.3.1.6 Drives and Affect

Some seek to keep agents benign through artificial feelings and needs, similarly to how animals modulate their prosocial tendencies [169, 32, 125]. For approaches that seek maximal biological plausibility, cognitive neuroscience offers significant bearing on the formulation and evaluation of ethics [178]. See section *Psychological Analogues* for exploration of additional potential similarities between types of intelligences.

4.3.2 Degree of Value Evolution

Even if an agent is able to be aligned accurately with some set of societally-acceptable values at a given point in time, for long-lived agents, one must consider, given that values evolve over time, how rigidly to adhere to those original values versus being open to updating those values. Determining whether, and how quickly, a long-lived agent can safely evolve, refine, or redefine the values it's initially imbued with remains an open problem [420, 284]. A hypothetical agent created hundreds of years ago, if created with strict value faithfulness and not allowed to update values, would exhibit what one today would consider very odd norms, values, and behavior [439]. See section *Ontology Update Thresholds* since mapping world events to any stale, ungrounded, or nolonger-founded concepts can lead to odd dynamics, and if those concepts are values the effects can be compounded.

4.3.3 Robust Human Imitation

One way to convey to artificial agents what one would like them to do is to show them in detail. Designing and training machine learning systems to effectively imitate humans who are engaged in complex and difficult tasks is an area of active research [437]. One important consideration in this approach is the set of tradeoffs between the depth of true objectives one would ideally like to uncover and the ease with which shorter-term goals or instrumental goals can be learned [80].

4.3.3.1 Inverse Reinforcement Learning

A system can infer the preferences of another rational, or nearly rational, actor by observing its behavior. A prominent family of algorithms for learning and imitating human behavior, as well as narrow or instrumental values, is inverse reinforcement learning (IRL) [437]. It attempts to learn the reward function that a human is approximately optimizing [365, 310]. Very related are apprenticeship learning [2, 1] and other methods for estimating a user's reward function [489, 351]. IRL has many variants that improve the generality, aggregability, or analysis of the reward function, or the workflow to approaching it [149]. A variant called interactive IRL attempts to learn about the reward function at the same time as trying to maximize it [24]. Issues with the unidentifiability of the reward function might also be addressable by active IRL [10]. IRL methods might not scale safely, however, due to their reliance on the faulty assumption that human demonstrators are optimizing for a specific reward function, where in reality humans are often irrational, ill-informed, incompetent, and immoral. Recent work has begun to address these issues [136, 137]. Generative adversarial imitation learning is another promising approach to this problem [205].

4.3.3.1.1 Cooperative Inverse Reinforcement Learning The cooperative inverse reinforcement learning paradigm views the human-agent interaction as a cooperative game where both players attempt to find a joint policy that maximizes the human's secret value function [185, 437]. An outstanding challenge in this is to determine which portions of the ascertained value function are instrumental, which are incidental, and which exhibit deep values. See section *Cooperative Inverse Reinforcement Learner* where this algorithm is used for control in an online context rather than upfront value learning.

4.3.3.2 Imitation Statistical Guarantees

Some techniques focus on having a measure of the robustness of human imitation within the contexts learned, such as those quantifying the sampling needed for preference inference [253, 437].

4.3.3.2.1 Generative Adversarial Networks Generative adversarial networks may be able to train a system to generate human-like answers to within some statistical confidence [172, 437].

4.3.3.2.2 Other Generative Models In cases of large output spaces and limited training data, one can use exploration, generation, and imitation variation to quantify confidence of imitation [181, 437], useful for imitation learning [226, 358, 27, 428].

4.3.3.3 Operator Modeling

In order to model an operator's preferences as comprehensively as possible, the operator themselves should be modeled well. By what methods can an operator be modeled in such a way that a model of the operator's preferences can not only be extracted, but also continually refined to become arbitrarily accurate, while representing the operator as a subsystem embedded within the larger world? [402] There is work on modeling individuals' cognition [174], progress in short-term modeling of, and adaptation to, operators [275], and modeling human preferences as inverse planning [37], but more work is needed to approach comprehensive modeling [403]. One approach to prevent overfitting is to use different assumptions about underlying cognitive models of the actor whose preferences are being learned [94].

For complex tasks, it seems plausible that the system will need to learn a detailed psychological model of a human if it is to imitate one, and that this might be significantly more difficult than training a system to do engineering directly. More research is needed to clarify whether imitation learning can scale efficiently to complex tasks. See section *Operator Value Learning*, which is similar but focuses on learning the ethical or moral values of people rather than their broader dynamics. See section *Psychological Analogues* as there will likely be much cross-pollination between these threads.

4.3.3.4 Reversible Tasks via Variational Autoencoding

For the subset of human tasks that are reversible, or able to be done either forwards or backwards with minimal information loss, one can form generative models based on training data of queries and associated good responses [437]. Variational autoencoders seem appropriate for this as they learn lower-dimensional manifolds that tease out conceptual structure [233]. It might also be possible to break such nonreversible tasks into multiple reversible tasks to leverage this technique [422]. Just which sets of tasks can be performed using reversible generative models remains underexplored.

4.3.3.5 Distance from User Demonstration

Another method of safely learning from observing humans is to learn baseline policies in an apprenticeship manner and explore with bounded deviations from demonstrations [358, 3, 11, 149]. It may however be difficult to learn deep preferences this way.

4.3.3.6 Narrow Value Learning

While the long-term goal of both value alignment and human imitation is to figure out the appropriate terminal or fundamental values, intermediate progress can also have appreciable utility for AI safety and beneficence. The learning of instrumental subobjectives, which may also be conceived of as narrow values [81] is a valid next step in value alignment. Recognition of the line between such instrumental values and fundamental values, however, is a significant outstanding need. See section *Inverse Reinforcement Learning* which provides many good methods for narrow value learning.

4.3.3.7 Scaling Judgement Learning

Generally, one can consider the combination of multiple learning and validation techniques for learning what a human would respond, or how they would judge a situation, reserving active learning and querying of humans for when it's quite necessary [86]. It has also been proposed that one might train a highly capable aligned agent using a mix of a series of more capable approval-directed reinforcement learning agents and bootstrapping [78]. In theory, it would be possible to establish a trajectory whereby the succession becomes capable of closely approximating what the human would have decided given much more time, resurces, and education. This is, however, much more difficult than developing a good generative model of observed human behavior. See section *Human Judgement Learner* where a similar technique is used in a continuous online manner.

4.4 Increasing Contextual Awareness

Whether using closed-world or open-world models, systems operating within the real world tend to model only a very small portion of the environment, rendering them oblivious to common sense contextualization of even concepts within their models. Such models also often have unintended structural biases and blind spots stemming from how they are generated, and reduction of this obliviousness leads to more trustworthy systems. With improved quality, salience, robustness, flexibility, and contextualization of conceptual models, agents can apply more common sense to introspection and to decision making. Typical ability gaps relative to humans include common sense context, learning causal models, the grounding of concepts, and learning to learn [249, 445], despite experimental systems having demonstrated each of these individually.

4.4.1 Realistic World-Models

Agents operating in a closed-world environment, e.g. AIs playing perfect information games like chess or imperfect information games like poker, only ever need to worry about the configuration of elements that they know exist. For agents embedded in the real world, however, invariably an open-world environment, awareness of the possibility of new or changing entities, relationships, and concepts needs to be integral to its processes of learning and pursuing goals.

4.4.1.1 Expressive Representations

The structures in which large open world models are modelled will need to be sufficiently expressive while remaining conducive to computational efficiency [370]. Other model desiderata include interpretability, paths to compatibility between separately trained systems, and paths to compatibility between symbolic and subsymbolic representations.

For highly capable systems, the content to fill out such models is too high-volume for humans to fill in directly, so either semisupervised, distantly supervised, or unsupervised learning methods are called for.

4.4.1.2 Unsupervised Model Learning

As real-world environments are very large, and supervised tagging of concepts and relationships does not scale, there is great interest in unsupervised model learning. In online contexts, explicit mechanisms are needed to deal with ongoing refactorings of the ontology, or world model, of a system [55]. See section *Degree of Value Evolution*, since as new concepts are learned and world models evolve, there are competing pressures to restructure existing values using the updated ontology and to change what set of values to hold to in the first place. See section *World Model Connectedness* as that addresses preventing fragmentation to benefit transfer learning and help avoid cognitive dissonance. The paradigm of maintaining a distribution of possible worlds and determining their likelihoods can be reasonable when those worlds collapse into a tractable number of classes [364].

4.4.1.2.1 Concept Drift Concept drift is the gradual warping of the meaning of a given concept over time. In a statistical learning context, the statistical properties of a target variable to be predicted, changes in unforeseen ways over time [158], and robust AI should be vigilant in detecting such drift. See section *Robustness to Distributional Shift*, which addresses identifying and responding to increases in ambiguity around a concept statistically.

4.4.1.2.2 Ontology Identification Given goals stated using a given ontology, formalism, or semi-formalism, and streams of data about the world, an advanced agent should be able to identify the most appropriate ontology in which satisfaction of the intended goals should be executed and evaluated [400, 403]. Doing this needs to account for the fact that those ever-changing streams of data are only indirectly and incompletely understood by the operator who originally provided the goals.

Relatedly, it would be helpful for the purpose to understand theoretically and practically how learned representations of high-level human concepts could be expected to generalize, or fail to do so, in radically new contexts [439].

4.4.1.2.3 Ontology Update Thresholds Ontology rot occurs when one or more concepts in a world model are held fixed as new dynamics emerge which merit updates to the world model [482, 286, 318]. When a measure or property becomes a target or a key performance indicator, the dynamics of the system change to optimize for it, often in unexpected or parochial ways, causing that metric to no longer be a good measure [284]. This is known as Goodhart's Law [288]. See section *Mild Optimization* because developing a formal model of Goodhart's Law can also benefit that need. Insufficiently adaptive ontologies can lead to parochial behaviors with justifications that may no longer be grounded [482, 318]. Ontologies updated too frequently or with too little evidence, however, can lead to mercurial behavior [285]. See section *Degree of Value Evolution* since, relatedly, when a property degrades in its role as a moral value, reevaluation of the intent of that value may be necessary.

4.4.1.2.4 Episodic Contexts In model-based reinforcement learning, one can use the observed transitions of the unlabeled episodes to improve the quality of the model [11, 18]. State abstraction from either online or offline episodes using similar such techniques can also be applied in other model learning paradigms [477].

4.4.1.3 Correlation of Dynamics

Taking into account correlations between the AI system's behaviors and those of its environment or of other agents is another kind of contextual awareness especially useful in game theoretic framings [441, 247, 204, 192, 405]. See section *Open Source Game Theory* as that addresses additional layers of possible agent correlation, stemming from logic.

4.4.1.4 World-Embedded Solomonoff Induction

There are theoretical challenges when the world model of an advanced agent includes itself, as more robust advanced agents would [403]. In addition to generally being more robust and accurate, this inclusion critically helps the agent include the effects of its internal dynamics on the environment. It can also provide a formal grounding for more realistic game theoretic analyses [258]. Updated paradigms are needed to model how generally intelligent agents that are embedded in their environments should reason [400, 328]. Identification of the induction problem itself, as analog to Solomonoff induction, in the scenario of an agent both embedded in and computed by its environment, is termed naturalized induction, and can benefit from more research [143, 400, 355]. The agent would need to consider what the possible environments that could be embedding the agent are, and determining a good simplicity prior would help to create a distribution or weighing of those possibilities [403].

Understanding this analog to Solomonoff Induction would help us circumscribe some theoretical limits around such agents. Deficiencies in such capabilities can lead to the type of agent that finds clever ways to seize control of its observation channel, rather than actually identifying and manipulating the features in the world that the reward function was intended to proxy for [400]. Determinations of how well an agent does in this regard would require formalizing that question in tandem with developing a scoring metric that evaluates the resulting environment histories, rather than just the agent's observations. Some theoretical progress has been made by using reflective oracles to model both agents and their environments, enabling agents to have models of one another and converge to various game theoretic equilibria [145]. Progress has also been made on the "grain of truth" problem, such that agents can reason using explicit models of each other without infinite regress [258]. Progress in this type of reasoning is also one of the prerequisites to the agent having a theory of open source game theory, useful for designing robust cooperation [113]. See section *Open Source Game Theory* as this type of model and reasoning has implications not only for directly improving contextual awareness, but can lead to foundational decision theoretic capabilities.

4.4.1.5 Perceptual Distance Agnostic World Models

Proximal world models centered on the agent's percepts [161] may lead to undesirable distortion of biases. When concepts or entities that are a shorter perceptual distance away accrue perceived value disproportionately due to that shorter distance, behaviors that seem narcissistic or overly parochial may emerge. These kinds of weaknesses can manifest when scaling classical reinforcement learners [128], for example. Indicative behaviors may include promotion of confirmation bias or incorrectly calculating externalities as negligible. A key preventative to such weaknesses is for world models, or at least significant portions of them, to be distal [161], establishing tenets or objective concepts as central. See section *Pursuing Environmental Goals* as that addresses the more general problem of objective function. See section *World Model Connectedness* as proximal and distal elements of an ontology should be richly connected for an agent to be robust. See section *Computational Humility* which can motivate the need for less agent-centric models.

4.4.1.6 Knowledge Representation Ensembles

One approach to improving contextual robustness is to maintain and to aggregate multiple different knowledge representation schemes, ontologies, structures, and biases [399]. See section *Multiobjective Optimization*, as a way to accomplish that is to have a different subobjective per knowledge representation scheme and using multiobjective optimization [284].

4.4.2 Endowing Common Sense

Common sense is often cited as a basic capability that humans have but AI and ML systems lack. Common sense is not a single faculty though, but a combination of extensive world knowledge, deep semantic and pragmatic parsing of percepts, and using the appropriate combination of types of reasoning for the situation [151, 271].

4.4.2.1 Common Sense Reasoning

Common sense reasoning includes using the appropriate combinations of induction, deduction, and abduction at the appropriate situations and times and using appropriate world knowledge [168], as well as dynamic binding to novel faculties [423].

4.4.2.2 Bootstrapped Common Sense

Abstracting patterns to establish semantics requires a dynamic mix of different reasoning abilities [452]. See section *Unsupervised Model Learning* which has many overlaps with this.

4.4.2.3 Seeded Common Sense

A core of common sense concepts, relationships, and dynamics can be used as as a pretraining or seeding to bootstrap a system that will learn additional ontology and knowledge around what is seeded [97]. When the core of the relevant contextual landscape can be specified via structured knowledge like this, one can effectively improve the system's contextual awareness by providing a higher quantity and broader scope of usable world-knowledge than one might expect the system to typically need. Abstracting patterns, which has similarities to unsupervised model learning, requires this kind of dynamic mix of reasoning abilities [452].

4.4.2.4 Metareasoning

Metareasoning, reasoning about reasoning, is an important common sense faculty that includes meta-level control of introspection and of computational activities and modelling the self [109, 103]. It helps to contextualize and modulate common sense reasoning and selection of appropriate algorithms [269].

4.4.2.5 Lengthening Horizons

Planning and lookahead time horizons significantly influence both the effectiveness of and the stability of agents. A potential technique to do so for online-planning markov decision processes is to do state abstraction [34]. See section *Follow-on Analysis.*

4.4.3 Concept Geometry

Conceptual spaces, be they vector spaces, semantic networks, more implicit subsymbolic structures, or symbolic-subsymbolic hybrid representations, are how concepts are defined, grounded, and related to each other [161]. Any ontology, including implicit ones in subsymbolic architectures, is biased by the faculties with which it was constructed and intended [411, 161]. See section *Value Structuring* as that addresses the geometry of human values. See section *Perceptual Distance Agnostic World Models* as that addresses proximal versus distal spaces.

4.4.3.1 Implicit Human Concepts

Beyond the realism of concepts, a possible desideratum for enhanced human compatibility is similarity as compared with how humans conceptualize the world. Concepts can be defined with respect to groundings to human-like primitives such as senses, maintaining human-like concept relationships and spaces [412, 445]. Hierarchical structured probabilistic models are one plausible avenue for human-level concept learning [248]. One should also aim to understand theoretically and practically how learned representations of high-level human concepts could be expected to generalize, or fail to do so, in radically new contexts [439].

4.4.3.2 Conservative Concepts

How can a classifier be trained to develop useful concepts that exclude quite atypical examples and edge cases? [437] Including negative examples to try to get a better sense of the classification boundaries is insufficient [171, 394]. Novel ways of analyzing cluster spaces may be necessary. See section *Generative Adversarial Networks* as actor-critic arrangements can refine the understanding of a concept, though more research needs to be done to scale the generalizations by the critic. See section *Goal Stability* as outlandish interpretations of concepts hamper stability. See section *Inductive Ambiguity Identification* as those techniques can be used to better analyze decision boundary neighborhoods and generated exemplars.

4.4.3.2.1 Dimensionality Reduction With Anomaly Detection One plausible and underexplored method of establishing conservative concepts is to find the important features of training instances, then use generative models to synthesize new examples that are similar to the training instances only with respect to those, and use anomaly detection to probe and repair the resulting space [437, 203].

4.4.3.3 Multilevel World-Models

Reality and possibility are very large. Multilevel conceptualizations therefore arise, whether implicitly or explicitly, in the layers of a deep net, in systems biology, in meta vs. upper vs. lower vs. operational ontologies, in variational renormalization in physics [270], and in multilevel analysis in statistics [398]. A key question for AI safety is how multi-level world-models can be constructed from sense data in a manner amenable to ontology identification [402, 403]. See section *Ontology Identification* for more on that.

4.4.3.4 World Model Connectedness

Fragmented world models occur when portions of an (explicit or implicit) ontology that, environmentally, should connect or overlap, remain cleaved. This can lead to limited causal learning, inconsistent theories, unwarranted concept splitting, and poor transfer learning [331, 284]. Fragmentation can also result in catastrophic inference, orphaning historical representations [314, 152]. Focus on establishing explicit connections across contexts to previously learned features might alleviate these issues [374].

4.4.4 Uncertainty Identification and Management

There are many types of uncertainty that AIs will need to model [403, 69]. As a matter of course, given their training data, machine learning deals with inductive uncertainty in typically narrow contexts. Awareness that insufficient traing data may have been provided for particular determinations supports the developing capability of recognizing ambiguity. Relatedly, determining dimensions or features for which there is little or no data, and which may be important, can be important for proactive ambiguity management [402]. This is particularly relevant in online learning, where training is always incomplete and actions may be taken to focus on improving areas of unclarity [390]. At any given point in time, there may properly be not only inductive uncertainty regarding empirical facts, but also logical uncertainty, being unsure of the as-yet-to-be-computed specific complex entailments of the things that are already known [163].

4.4.4.1 Inferring Latent Variables

One can Identify latent or implicit variables from data via dependency analyses on that data [437]. See section *Unsupervised Model Learning* as finding these is a key ability for that.

4.4.4.2 Inductive Ambiguity Identification

A desirable general capability for machine learning systems would be to detect and notify us of cases where classifications of test data is highly underdetermined given the training data [437]. To aide such robustness, a system should be able to characterize the probability of disagreement or local ambiguity around a concept in a concept space [194]. See section *Adversarial ML* as that also deals with robustness to changes in input distributions, introduced maliciously and with potentially worst-case scenarios.

4.4.4.2.1 Scalable Oversight of Ambiguities Methods for efficiently scaling up the ability of human overseers to supervise ML systems in scenarios where human feedback is expensive, thus clever strategies are needed for making most efficient use of such resources. See section *Scalable Oversight* which describes a scalable disintermediated active learning.

4.4.4.2.2 Bayesian Feature Selection Given sufficient data, identifying and using the right features in order to model uncertainty well can be approached with bayesian structural analyses of the data [276, 182, 437].

4.4.4.2.3 Non-Bayesian Confidence Estimation Most machine learning methods do not by default identify ambiguities well, either lacking such a concept entirely, or often poorly calibrated at doing so even if overall accuracy is high [437]. It is well known that neural networks are often overconfident in their results, but there has been recent work to calibrate their confidences better [171, 311]. The techniques of conformal prediction attempt to produce well-calibrated sets of possible predictions [457].

4.4.4.2.4 Active Learning Judiciously asking humans to label appropriate examples in order to disambiguate among hypotheses, active learning, can be used when ambiguities are detected at a scale or speed where querying humans is practical [387, 195, 52, 388, 51, 79, 437].

4.4.4.2.5 Realistic-Prior Design If there is prior knowledge of a domain, bayesian priors can be designed especially to reflect its actual causal structure, reducing ambiguities [101, 437], and this may be automated. Knowledge-informed expectation propagation can likewise be formulated in non-bayesian formats [469].

4.4.4.2.6 Knows-What-It-Knows Learning A thin layer of metaknowledge within an ensemble can provide an easy way to detect ambiguities. Knows-What-It-Knows learning maintains set of plausible hypotheses and when there is disagreement among components, outputs an Unclear state, prompting finite queries to humans to provide golden data [437, 263, 231, 430].

4.4.4.2.7 Robustness to Distributional Shift For general robustness of a learning system, one aims to get a system trained on one distribution to perform reasonably when deployed under a different distribution [11]. Deep nets can even incorrectly output very different outputs for only trivially different inputs [429]. Where applicable, a potential solution is to train on additional subdistributions or variations of a distribution in order to have a more robust model [12], but often the nature, direction, or breadth of the drift will be unknown in advance. There are, however, some approaches to learning with concept drift [199, 158] in an online context. Multiple research areas relevant to this capability, including change detection, anomaly detection, hypothesis testing, transfer learning, and others [11]. A variety of different statistical techniques can be employed to detect or mitigate these issues [11]. The method of moments, characterizing learned distributions via their top moments, is a partially-specified model technique for assessing instrumental variables [13, 197, 11]. It enables unsupervised learning, including estimating latent variables, under conditional independence assumptions [196, 346]. For well-specified models, online-learning and testing of generative conditional independence structures is another promising avenue for distributional shift robustness [292, 313, 11, 110, 268, 265].

4.4.4.2.7.1 Covariate Shift When the distribution of the inputs used as predictors changes between the training and the testing or production stages, covariate shift analysis can inform as to the extent [11]. For well-specified models, online-learning retunings to parameters or to sample weights can be appropriate [11]. For highly expressive model families, though some work has been done, there still needs more exploration to see how well they can predict their out-of-sample performance [206, 408, 409, 177, 227, 329, 50]. The sample selection bias may also be able to be leveraged to this end [77].

4.4.4.2.7.2 Unsupervised Risk Estimation Given a model and unlabeled data from a test distribution, unsupervised risk estimation [134, 417, 418, 121, 487, 38] can aide estimating the labeled risk of the model [342, 11, 218].

4.4.4.2.7.3 Causal Identification Techniques from econometrics can be used to estimate causal structure and instrumental variables from data [11, 378, 379]. See section *Causal Accounting* which would be a consumer of such determinations.

4.4.4.2.7.4 Limited-Information Maximum Likelihood Limited-information maximum likelihood is a technique from econometrics for accomodating partially specified models, and partial specification allows for broader and more robust coverage of test distributions [11, 16, 17].

4.4.4.2.7.5 Active Calibration Active calibration is a technique that helps to pinpoint the structural aspects of most uncertainty within a model [11, 463, 231]. It is a way of obtaining calibration in structured output settings, as a followup action to recognizing inputs being out-of-distribution [242].

4.4.4.2.7.6 Reachability Analysis Reachability analysis is a technique for when a system model is known, for bounding controls and disturbances within which the system is guaranteed to remain safe [279, 295]. This aides alleviating situational ambiguity [11].

4.4.4.2.7.7 Robust Policy Improvement Once ambiguities are recognized, a potential response is robust policy improvement, finding a policy that minimizes the downside, or alternatively maximizes the utility of the worst case scenario, over that set of ambiguities or distribution of uncertainties [464, 11].

4.4.4.2.7.8 Counterfactual Reasoning Consideration and reasoning about counterfactual scenarios, or what would have happened if the world were different in a certain way [308, 362, 335], based on logical entailment but applied in a machine learning context [62, 340, 11, 427, 225, 391] enables more accurate, comprehensive, and aligned enumeration and evaluation of options and causal learning. It may be possible to use Garrabrant inductors to predict counterfactuals [45]. See section *Theory of Counterfactuals* as there are ongoing theoretical developments in how possibilities should be enumerated and modeled. See section *Decision Theory* for a broader treatment of counterfactuals oriented at goal stability. See section *Logical Counterfactuals* which aims to progress methods for enumerating them.

4.4.4.2.7.9 ML with Contracts To minimize ambiguities at design time, one can construct machine learning systems that satisfy a well-defined contract on their behavior, like in software design [267, 61, 11, 384]. See section *Verified Component Design Approaches* which can benefit from these techniques.

4.4.4.2.7.10 Model Repair Another potential response to ambiguity identification is model repair, altering the trained model to ensure that certain desired safety properties will still hold [166, 11].

4.4.4.3 Resource-Aware Reasoning

Given unbounded time and memory, an AI system would be able to figure out both the full distribution of reasonable inferences and the full set of logical entailments that stem from the facts it already knows. Realistic agents, however, have time and memory constraints, and must manage those resources in their activities, including their computations [191]. See section *Logical Uncertainty* as that foundational theoretical topic, which is still evolving, really underlies much of this topic to a degree not traditionally acknowledged.

4.4.4.3.1 Decision Under Bounded Computational Resources Properly budgeting and prioritizing finite resources in order to accomplish any other goals will lead to more robust and higher quality actions [371]. Relatively applied explorations of reasoning and decision making under bounded computational resources have been done [207, 363], but there is also more theoretical progress needed [104, 103]. One class of approaches prominently considers correlations between the bounded agent and its environment [191]. More explicit treatments of practical decision theory under such agents also follows from this [192]. Ongoing progress in unifying structured logic and probability for generalized empirical inference [370] will aid this capability. See section *Metareasoning* as abstract consideration of logic flow can be used to modulate tactics and resource usage.

4.4.4.3.2 Logical Induction Very large world models, time bounded decision making requirements, high branching factors, and generally large deterministic computations can lead to appreciable periods when logical entailments of known facts are not yet known. Many classical models of agents, such as in game theory, economics, and bayesian reasoning, assume such periods are inconsequential, that such agents are logically omniscient, but there is a growing realization that advanced AI agents will need an explicit model of these unknown knowns. Indeed, nearly all practical reasoning implicitly involves logical uncertainty [155]. The problem of an agent reasoning about an environment in which it is embedded as a subprocess is actually inherently a problem of reasoning under logical uncertainty [403].

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 [403]. This would model confidence in assertions and how those confidences change over time with additional reasoning. Significant theoretical progress has recently been made in logical induction, including Garrabrant inductors [163] and optimal polynomial-time estimators [240], which bridges rational choice theory and probability theory, combining logic and observation and their respective uncertainties. Much more work is required on this topic to develop it to practicality. It was shown that, in the limit, such induction can outpace deduction [164]. Logical induction also facilitates alternating between different types of logic, as is particularly needed in open-world domains.

4.4.4.3.2.1 Logical Priors The question of how to arrive at a reasonable prior for a logical sentence presents itself when starting induction about logic [404]. While providing good starting points, it may actually be much less important in the limit [164]. In order to address arbitrarily deep combinations of logical uncertainties, however, additional research is needed on what satisfactory set of priors over logical sentences a bounded reasoner can approximate [126, 87, 403] in practice.

4.4.4.3.2.2 Impossible Possibilities Exploring how deductively limited reasoners can perform counterpossible reasoning, or counterfactual reasoning from impossible antecedents [49], will help progress theory and usage of both counterfactuals and logical induction [403]. This is equivalent to asking how deductively limited reasoners can approximate reasoning according to a probability distribution on complete theories of logic [87]. Early theoretical work has considered counterpossible reasoning in deterministic settings, but further development can aid agent theory and decision theoretic foundations [406].

4.4.5 Symbolic-Subsymbolic Integration

Integration between symbolic, or explicit discrete models, and subsymbolic, or implicit numerical models [72, 74, 334, 6], would enable meaning and value to propagate between different architectures or modalities more fluidly [162, 453, 466, 169, chap. 9], and would make conglomerations of different AI and ML components in agents more robust. This integration facilitates the grounding of concepts as well, since sensory data, sensory processing, and actuation processing all have significant subsymbolic components [162, 445].

4.4.5.1 Use of Structured Knowledge In Subsymbolic Systems

The meaningful use of symbolic concepts, relationships, constraints, and rules in subsymbolic, typically statistical, e.g. neural net, contexts can help orient deep learning based agents with concepts and relationships its operators care about [70, 162, 453, 334].

4.4.5.2 Use of Subsymbolic Processing In Symbolic Systems

Systems that aggregate and integrate disparate components using explicit semantic interlingua can use subsymbolic processing for determining what structures, statements, and connections should exist and how they should be weighted [198], and can do novel methods of reasoning over them [118]. Though symbolic-first systems have become less popular of late, they are more amenable to integration with the systems that already run society. See section *Monitoring* as such architectures can benefit interpretability.

4.5 Psychological Analogues

Artificial intelligences, including organizations and AI agents, and natural intelligences, including animals and humans, share a number of features in common [425]. Information processing, knowledge processing, planning, and action are at the core of all of these classes of intelligent agent. In many cases, even when architectures are appreciably different, similar dynamics can occur, both in the proper and improper processing of knowledge [174, 426]. Though interesting, what motivates cross-modality analysis is that establishing correlates, analogues, or metaphors that run deep between natural cognition and artificial cognition might enable us to harvest and transfer insights in both directions between psychology (even organizational psychology and animal psychology) and artificial intelligence safety research. See section *Operator Modeling* which attempts to model humans to understand what they want and what they'll do, a more direct safety application of psychological modeling.

4.5.1 Cognitive Parallels

One can more broadly analyze human cognition and developmental psychology in terms of machine learning algorithms and vice versa to potentially elucidate fruitful avenues of research [36, 174, 221, 440, 220]. In so doing, the field should want to establish which general classes of analogues can be valid, for which architectures, and to what extent [426], and this can benefit from additional research. See section *Drives and Affect* as regards attempts at cognitive-inspired prosocial mechanisms.

4.5.2 Developmental Psychology

Within the field of artificial general intelligence, and with some precursor machine learning techniques, parallels are drawn between human early childhood learning and development and seed artificial cognitive systems that are either grown or learned [420, 220, 440, 169]. Questions remain open as to how far this metaphor extends, particularly with respect to architectures that are not intended to be biologically plausible. See section *Whole Brain Emulation Safety* as questions also remain open as to how far such metaphors extend even with architectures that are meant to be biologically accurate.

4.5.3 Dysfunctional Systems of Thought

Within each corresponding pair, generalizing some key dynamics of the problem, and optionally some subset of the treatment for the psychological condition, may lead to insights applicable to such issues with artificial agents. A particularly promising application of analyzing these natural-artificial analogues is to mine insights about analogous issues from each domain that may be applicable to the other domain [31]. Indeed many of the dysfunctions that can affect naturally developed minds have close counterparts in synthetic intelligences, and vice versa [28, 284]. By characterizing these correspondences more specifically [426, 354, 314, 152, 374], researchers might be able to find insights from the field of psychology and apply them to AI, or, in the interest of furthering science, helping people, and strengthening bridges across these fields, perhaps find insights from AI safety and translate them into insights psychologists can use.

4.5.4 Whole Brain Emulation Safety

Psychological analogues to AI are relevant not only for engineered AI but also for scanned and emulated biological brains. Though neuroscientists understand much of what occurs in such detailed and bioplausible neural networks, that understanding is at the level of annotations rather than being able to produce a detailed explanatory or generative model. This may lead us to treat the emulation as a black box, and so psychology-derived AI safety approaches and characterizations would become quite relevant [377].

4.6 Testing and Quality Assurance

Traditional software quality assurance techniques will not scale well with highly capable AI, but they can still be deployed to provide early and efficient warnings about robustness or safety issues, and analogues to these techniques can be helpful. Unit tests, for instance, can perform some basic checks of sanity and deception. To extend such tests toward the realm of general-purpose agents, a task theory and framework is necessary [446]. These analogues will not be nearly sufficient to maintain safety in the traditional quality control paradigm, hence the other safety techniques here [403].

5 Security

There are a wide variety of ways AI can go wrong ; many of them involve novel endogenous risks, but many of them also involve bad faith corruption or attacks at any of a multitude of levels [473].

5.1 Standard IT Security

Cybersecurity is applicable and important to any kind of software, and artificial intelligence is no exception [371]. AI failures can come about through lapses in security [471, 40]. Unique to AI are the specific vulnerable or sensitive targets of attack within an AI system, and the active ways of defending those [474, 332].

5.1.1 Verified Downstack Software

Because software verification and correct-by-construction are such powerful techniques to reduce the kinds of implementation issues that are often exploited during security events, utilizing a stack of as much verified software, as high as is practical in the stack, would be prudent. Even just very well-tested underlying platform can provide similar benefits, e.g. the DARPA SAFE program aims to build an integrated hardware-software system with a flexible metadata rule engine [123], on which can be built memory safety, fault isolation, and other protocols that could improve security by preventing exploitable flaws. Also see relevant sections *Formal Software Verification*, *Verified Component Design Approaches*, and *Careful Engineering.*

5.1.2 Utility Function Security

It would make sense to establish extra privileges and protections around modifications of an explicit or implicit utility function of an AI, given that it's a highly centralized and leveraged point [474].

5.1.3 AI to Support Security

For advanced and general purpose agents, it should be quite possible to utilize its own advanced automation and intelligence analysis capabilities to, analyze potential malware [353], protect itself against active exploits, detect intrusions [252], and automatically check for potential exploits it may be vulnerable to. As AI matures, AIs one wishes to keep good will also need to protect themselves against not just human hackers, but also AIs that have either become, or were designed to be, bad [341]. Likewise, both context-aware and narrow AIs will need to protect the security and privacy of others they are serving, so an understanding and application of differential privacy [488, 447] practices would be called for [371].

5.1.4 Security-Validated Software Engineering

When creating advanced, powerful, or high-stakes AI systems, security-validated software engineering practices are called for. Given that vulnerabilities can use a growing array of unexpected side-channels [184], and given that advanced agents can be more creative about side-channel use than people are, a side-channel vulnerability assessment would be important to such validation, yet it wouldn't be able to provide any guarantees.

5.2 Red Team Analysis

Red team analysis uses an adversarial entity or team to probe and identify vulnerabilities in the system of interest, usually in a blackbox context. The recent DARPA Cyber Grand Challenge likewise had various automated adversarial teams whitebox scan their own systems and blackbox scan their opponents, patch themselves, develop exploits, and penetrate opponent systems, all in a matter of minutes rather than the weeks to months that humans would normally take for the same steps [410, 306]. See section *Automated Vulnerability Finding*, which addresses finding vulnerabilities in whiteboxes rather than blackboxes.

5.2.1 Penetration Testing

Penetration testing specifically uses an adversarial entity to attempt to break into the system of interest in various ways and to go as deeply into its systems as possible. Agents can analyze the vulnerabilities of other systems and agents in this way [66], and can correlate this analysis with analyses of the likely source code generating those vulnerabilities.

5.3 Tripwires

To catch internal misbehavior, one can introduce plausible vulnerabilities, with alert triggers, that an agent can exploit and will only do so if its reward function is being gamed [30, 11]. This technique is also useful for catching intrusions by external threats as it is classically applied. Also see relevant sections *Avoiding Reward Hacking* and *Buried Honeypot Avoidance.*

5.4 Containment

Since generally intelligent agents represent varied and storied risks, it is prudent to test such agents in a confined environment [371] before releasing them into the wild. Building and using containers where such tests on powerful agents can be done safely and reliably [476, 30] is challenging, and must account for operator psychology. If general containment proves too difficult, it may be wise to designing the AI and a specialized container for it in parallel [58].

5.5 Handling Improper External Behavior

Third parties, including other AIs and humans, may be malevolent [341], adversarial, or incompetent, and there are some techniques to detect and protect against them.

5.5.1 Adversarial ML

An agent is more robust when one can guarantee its good behavior even when an adversary picks test or production data for it, from a different distribution than training data, aiming to make the agent fail [437]. It is increasingly common that machine learning systems can be manipulated [332, 171, 300], and this can even happen with real-world examples [245]. See section *Inductive Ambiguity Identification* since dynamics-induced ambiguity issues and adversarially-introduced ones share many features, and many of the techniques that help with the former will also help with the latter. See section *Counterexample Resistance* as this issue is core to concept learning and validation as well.

5.5.1.1 Dealing with Adversarial Testing

In online scenarios where actual adversaries may provide very skewed or poisoned data, it can be imperative that the agent not be corrupted by that [212, 307]. There is a growing literature on learning in a manner that is robust to such adversaries [212]. See section *Generative Adversarial Networks* for contrast, since adversarial machine learning as referenced deals with external adversaries, as where GANs as referenced contain the adversarial structure as an internal and constructive critic meant to improve its performance. See section *Sensibility-Triggered Defense.*

5.5.2 Statistical-Behavioral Trust Establishment

In order to determine which sources of information, or which agents, in an environment can be trusted, techniques that analyze those entities' behaviors statistically may be employed [371]. Intelligently managing trust in this manner can be applied to other AI systems [348], and generally, reputation models [375] allow such mechanisms to scale [307].

5.5.3 Modeling Operator Intent

In use cases where it is unclear which subset of operators [307] to be loyal to [341], it may be proper to model the intent of each operator [348] and bias to those that reflect values [360] present in the agent [420] after successful value alignment [402].

5.5.4 Sensibility-Triggered Defense

When a privileged bias is challenged by pressure from particular data, that data source should be flagged as suspect and trusted less, at least until such time that there is more information about the case [307, 433, 42, 431].

5.5.5 Detecting and Managing Low Competence

Some third-party agents may initially appear malicious but may actually be low-competence and not malicious, and a robust agent should detect and manage that [419, 431].

5.6 Privileged Biases

Specific biases, relationship valences, or soft or hard constraints within an agent's world model can be annotated as privileged and weighted highly in tradeoff calculations [433]. These may model values. Such biases can be allowed narrower tolerances, allowed to change more slowly, or have associated triggers informing the agent of adversarial pressures or attacks on them [431, 42, 266]. See section *Structured and Privileged Ethical Biases*.

5.7 Norm Denial of Service

Soft constraints such as norms may have associated recovery mechanisms that can be susceptible to timing, concurrency, and volume-based exploitation. Denial of service attacks on environmentally-enforced recovery or safety mechanisms may therefore occur when an agent launches a sustained volley of violations against such environmental soft constraints [71, 278]. If recovery mechanisms lack a meta level flow control to guard against such situations and are unable to meet the concurrency demanded, such an attack might cause recovery procedures associated with subsequent violations to be incompletely processed or uninitiated [175, 278]. See section *Grounded Ethical Evolution* which may be susceptible to this. See section *Metareasoning* which might seem called for in constraint enforcement mechanisms. See section *Logical Induction* which can in theory be used in service of meta level flow in such situations. See section *Handling Improper External Behavior* as this can also be regarding third-party entities.

5.8 Misuse Risk

Another type of security risk is that a highly capable agent can be misused by its operators [473]. There may well be technical strategies, e.g. requiring a large number of operators or sources of values, to help mitigate this risk, but this is very underexplored. See section *Modeling Operator Intent* which addresses detection of discord within controlling operators with respect to each other and with respect to previously loaded values.

6 Control

It is often desirable to retain some form of meaningful human control [371], whether this means a human in the loop or on the loop [200, 333], yet the system should have a clear expectation of whether or not the human is sufficiently experienced, skilled, and ready for what it will ask of them [319]. It has been argued that very general, capable, autonomous AI systems will often be subject to effects that increase the difficulty of maintaining meaningful human control [322, 59, 58, 392].

For advanced agents, most goals would actually put the agent at odds with human interests by default, giving it incentives to deceive or manipulate its human operators and to resist interventions designed to change or debug its behavior [58]. This is because if an AI system is selecting the actions that best allow it to complete a given task, then avoiding conditions that prevent the system from continuing to pursue the task is a natural consequence and can manifest as an emergent subgoal [322, 59]. That could become problematic, however, if one wishes to repurpose the system, to deactivate it, or to significantly alter its decision-making process; such a system would be rational to avoid these changes. Following a broader convention and understanding within the artificial intelligence and computer science domains, self-control and reliable decision making fall under validation while operator control falls in the present section. See section *Averting Instrumental Incentives* therefore, as that contains many of the key prerequisites to reliable control by operators, specifically reliable self-control. If methods of alignment and control don't scale with how the intelligence of the system scales, a widening gap of potentially dangerous behavior will appear [89]. Advanced systems that do not have such issues are termed corrigible systems [407].

The possibility of rapid, sustained self-improvement has been highlighted by past and current projects on the future of AI as potentially valuable to the project of maintaining reliable control in the long term. Technical work would likely lead to enhanced understanding of the likelihood of such phenomena, and the nature, risks, and overall outcomes associated with different conceived variants [209, 208]. Theoretical and forecasting work on intelligence explosion and superintelligence have been done before, but require regular updating and improved methods [73, 58]. Yet other architectures seek situations that seem unconstrained within some given time horizon [468], which present additional control challenges. There will be technical work needed in order to ensure that meaningful human control is maintained for the variety of critical applications [131] and architectures.

6.1 Computational Deference

Researchers argue that powerful artificial agents should choose to respect and submit themselves to their operators. When there is a tight coupling in such AI-operator systems, this requires factoring the human's behavior in more [9] e.g. issues with alerting non-alert user to take over driving self-driving car [319]. See section *Computational Humility* as limiting an agent's estimation of its own importance can be an important prerequisite to holding its operators in relatively high esteem. Another type of deference would be in service of stabilizing or averting activities that would noticably result in the environment changing to accomodate the agent [193]. Those activities are another manifestation of Goodhart's Law [288]. See section *Mild Optimization* because developing a formal model of Goodhart's Law can also benefit that need.

6.1.1 Corrigibility

Another natural subgoal for AI systems pursuing a given goal is the acquisition of resources of a variety of kinds. For example, information about the environment, safety from disruption, and improved freedom of action (such as by additional compute power) are all instrumentally useful for many tasks [322, 59]. By default, an advanced agent has incentives to preserve its own preferences, even if those happen to conflict with the actual intentions of its developers [321]. Uncommon kinds of reasoning may be required to reflect the fact that an agent is incomplete and potentially flawed in dangerous ways [404, 403]. For example, the agent may need to consider counterfactuals for each effect of an action in order to finally ignore the effects of a given channel [224]. There may also be a path to a similar effect via differentially private multiarmed bandits (a prioritization mechanism where private information is connected to individual rewards) [447]. See section *Consistent Decision Making* as those considerations, e.g. goal stability, are formidable prerequisites to corrigibility.

6.1.1.1 Interruptability

Interruptability, being able to have a powerful agent safely and unbiasedly accept interruption or shut down commands by an operator, may be aided by having the system include control information about itself when modeling what its operator wants [186].

6.1.2 Utility Indifference

A powerful AI should be structurally indifferent to having its objective function switched out by valid operators [403]. An ongoing area of research is how a utility function can be specified such that agents maximizing that utility function switch their preferences on demand, without having incentives to cause or prevent the switching [22, 20].

6.1.2.1 Switchable Objective Functions

Combining objective functions in such a way that the humans have the ability to switch which of those functions an agent is optimizing, but such that the agent does not have incentives to cause or prevent this switch [407, 22, 326], would be a key capability [437].

6.1.3 Control Transfer

Identifying situations where control should be transferred, both to and away from humans, is necessary for a broad range of active learning, deference, cooperative learning, and low-confidence situations [9], applicable to both short and long timeframes [371].

6.2 Oversight

Oversight techniques aim to enable operators to effectively monitor, direct, and control an advanced agent, and give them the knowledge they need to do so in an informed manner. These techniques can be largely applicable in the case when the operator is another artificial agent as well, given sufficiently advanced AI. Researchers are driven to ask questions like how one might train a reinforcement learning system to take actions that aid an intelligent overseer, such as a human, in accurately assessing the system's performance [437]. This exposes challenges such as supervising machine learning systems in scenarios where they are complex and even potentially deceptive. See section *Value Learning* which shares many common techniques and dynamics with oversight.

6.2.1 Scalable Oversight

To scale oversight, it is desirable to ensure safe behavior of an agent even if given only limited access to its true objective function [89]. Methods for efficiently scaling up the ability of human overseers to supervise machine learning systems in scenarios where human feedback is expensive are promising, if early [11]. See section *Scaling Judgement Learning* in which such scale is also addressed in a more upfront or potentially offline manner. One potential approach is semisupervised reinforcement learning, where the agent sees the reward signal for only a small subset of steps or trials [11, 90]. See section *Unsupervised Model Learning*.

6.2.1.1 Supervised Reward Learning

With a supervised reward learning approach, one can train a model to predict the reward from the state on either a per-timestep or per-episode basis, and use it to estimate the payoff of unlabelled episodes [11], with some appropriate weighting or uncertainty estimate to account for lower confidence in estimated versus known reward [117, 382].

6.2.1.2 Semisupervised Reward Learning

A somewhat scalable approach is to train a reinforcement learning model in a semi-supervised manner [148] to predict the reward from the state on either a per-timestep or per-episode basis, and use it to estimate the payoff [11] of unlabelled episodes, with some appropriate weighting or uncertainty estimate to account for lower confidence in estimated vs known reward [117].

6.2.1.3 Active Reward Learning

With active reward learning, one trains a model in an active learning manner to predict the reward from the state on either a per-timestep or per-episode basis, and use it to estimate the payoff of unlabelled episodes [117], e.g. identifying salient events in the environment and querying human operators as to the respective rewards [11].

6.2.1.4 Unsupervised Value Iteration

The technique of unsupervised value iteration may also be useful for scalable oversight. This involves using the observed transitions within unlabeled episodes to make more accurate Bellman updates in the context of a markov decision process [11].

6.2.1.5 Distant Supervision

Distant supervision in this context would be where humans provide some useful information about the system's decisions in the aggregate, or some noisy hints about the correct evaluations [11], like some techniques within weakly supervised learning [290, 135, 393, 183, 167, 294] Broad paradigms like reward engineering [88, 129] should also be considered in such a context to avert unhelpful instrumental incentives.

6.2.1.6 Hierarchical Reinforcement Learning

Hierarchical reinforcement learning offers a compelling technique for scaling delegation and oversight [11]. In this approach, there is a high-level RL agent, operating on abstract strategies, receiving likely-sparse reward signal feedback from above it, and having it delegate to lower-level more object level RL agents, for which it in turn generates synthetic reward signals, with many such levels [122, 244]. This formulation is somewhat of a microcosm of AI safety overall, since subagents can do things that don't necessarily serve its higher-level agent's goals [11, 257]. Such a structure may prove amenable to fostering situational awareness via risk-conscious skills, which would strengthen its utility for the present task [289].

6.2.1.7 Trusted Policy Oversight

Given a trusted policy, an agent can be made to explore only regions of state space that the policy strongly believes can be recovered from [11]. See section *Bounded Exploration* which uses a very similar technique specifically for safe exploration.

6.2.1.8 Cooperative Inverse Reinforcement Learner

The cooperative inverse reinforcement learning paradigm views the human-agent interaction as a cooperative game where both players attempt to find a joint policy that maximizes the human's secret value function [185, 437]. An outstanding challenge in this is to determine which portions of the ascertained value function are instrumental, which are incidental, and which exhibit deep values. See section *Cooperative Inverse Reinforcement Learner* where this algorithm is used for more upfront value or preference learning.

6.2.1.9 Human Judgement Learner

Techniques for learning human judgement are another approach to scalable control [437]. While some are less scalable or more narrow [237], others aim to be highly scalable [82]. In the latter case, one might train a reinforcement learning system to take actions that a human would rate highly by using a framework where the system has to learn the "human judgment" reward function, and where training data is produced by having a more advanced agent, e.g. a human, evaluate the learner's actions [86], approval-directed agents [78].

In ambitious conceptions of this, the goal is not just to form a good generative model of observed human judgement, but the much more difficult goal of using the trajectories of past learning to extrapolate forward to what the subject would decide given more time, education, and resources. See section *Inverse Reinforcement Learning*, a practicable technique for apprenticeship learning and determining human goals. See section *Scaling Judgement Learning* where a comparable technique is used for upfront value learning.

6.2.1.10 Informed Oversight

In oversight scenarios like the approval directed agent paradigm, the more powerful and well-informed principal agent will need to find ways to incentivize good behavior from the weaker and more ignorant agent that the more capable one would like to teach [92, 223]. Underexplored as yet are the problems of informed oversight that come about when the system is highly capable and might be able to manipulate its human supervisors or circumvent their efforts [437]. Determining what sort of guarantees one should want in order to justify confidence in their ability to assess a system's behavior in the first place would be useful to provide next research steps on the theoretical side [437]. See section *Monitoring* which provides key ways for the operator to be informed. See section *Transparent Reinforcement Learners* as sufficiently transparent RL might open practical approaches to informed oversight [92]. See section *Value Alignment* which, to scale, may require viable informed oversight [92].

6.2.2 Controlling Another Algorithm

As time progresses and agents become sufficiently advanced, automatic generation of subagents may well occur. While trivial with simple fixed algorithms, there is currently extremely little research on the general case of how one algorithm can control another one with potentially independent optimization criteria [112]. See section *Open Source Game Theory* which addresses this issue in the case that source code of both agents are known by both.

6.2.3 Capability Distillation

Being able to distill or crystalize the capabilities, knowledge [202], and skills [228] of one system into another more compact one, or a less capable one, at least to within some tolerance, would be useful for many of the approaches to oversight and control. For applicability to long-term AI safety, being able to learn these in the face of much scarcer feedback containing possible inconsistencies would seem to be necessary and requires more research [83].

6.2.4 Rich Understanding of Human Commands

Though some advanced approaches aim to put the full burden of diambiguating humans on the intelligent agent, methods for effective, robust, meaningful, and minimally ambiguous communication between humans and machines [115] may well be necessary. It has been argued, however, that agents with different world models and only symbolic serialization between face very high computational complexity when statically attempting exhaustive disambiguation [211, 216]. Also see relevant sections *Increasing Contextual Awareness* and *Implicit Human Concepts.*

6.2.5 Monitoring

Monitoring, such as transparency or interpretability, can, to a degree [273], help humans or other agents understand what an agent is doing, why it is doing it, and what alternatives it has considered [302]. See section *Symbolic-Subsymbolic Integration* because conceptual structure can be used to improve machine learning interpretability [100].

6.2.5.1 Transparency of Whiteboxes

Whitebox algorithms, which are amenable to introspection, each have respective methods by which algorithms can be more transparent to introspection and understandability [437], including more explicitly interpretable models such as bayesian networks, causal networks, and rule lists, [153, 337, 461, 72, 219, 261, 108, 157] and even less typically-interpretable algorithms such as graphical models and dimensionality reduction [454, 280, 100]. Techniques like listing an agent's most likely next actions, [443, 262] summarizing its upcoming possible future states, [34] interactive behavior introspection and exploration, [238, 395, 243, 239, 251, 64] and generating narrative explanations of decisions [53] all support this type of transparency, and can potentially be used in conjunction. The understanding and visualization of representations learned by neural networks is both a large need for the increasingly popular and powerful deep learning techniques and has a growing set of proposed solutions [397, 484, 301, 283, 171, 230, 241, 320, 312, 70]. When trying to understand complex deep learning systems, generating visualizations or exemplars that accentuate particular parts of a the deep network that are particularly relevant for the classification [397, 484] can be useful.

6.2.5.1.1 Transparent Reinforcement Learners Techniques for analyzing and reporting on the models and policies of agents learned through reinforcement learning [483, 44] will have growing import as reinforcement learning becomes more widespread. See section *Informed Oversight* as transparent RL may help a lot with this core control problem [93].

6.2.5.2 Interpretability of Blackboxes

Blackbox interpretability is qualitatively different from whitebox transparency in that individual blackbox decisions are by definition inscrutible. The reporting or visualization of how decisions would have been different given different variations of the inputs, or reporting which of the input's features were most important in making a decision, allows an operator some understanding [450, 437] of models in a coarse-grain manner. One can also define coarse abstractions of neural networks that can be more easily verified to satisfy safety constraints [349]. Methods for explaining classifications that finds a sparse linear approximation to the local decision boundary of a given black-box ML system [448, 352, 291] may also be useful. Similarly, some techniques can report the gradient in the input of the classification judgment, [33] which can be used for exploratory heatmap visualizations of expected behavior. Metrics for reporting the influence of various inputs on the output of a black-box ML system [119, 356, 450] might be the simplest way to report on why blackbox or proprietary algorithms have reached particular decisions. For deeper explanations, one may consider training systems to output both an answer, such as an action, and a report intended to help the operator evaluate that answer [93].

6.2.5.3 Adversarial Transparency

When systems are made transparent, or at least seemingly interpretable, novel security issues can arise. The objective function or the explanatory function may become gameable [165]. See section *Adversarial ML* as adversaries may take advantage of transparency to exacerbate such challenges. See section *Open*

Source Game Theory because radical transparency in the contexts of creating trustable subagents or successors, or of control of one algorithm by another, can actually lead to much better outcomes for all parties.

6.2.5.4 Visualization and Situational Awareness

An understanding of the context and other salient operational information is key for a human operator monitoring the state of an AI. Methods to provide such situational awareness as completely and efficiently as possible would aid monitoring significantly. Situational awareness includes the broader context as the AI sees it, a representation of the environment, auto-selected salient factors for decision making and their recent and projected values, characterizations of risk sensitivity, and the top reasons for both recent decisions and upcoming decisions [259].

6.2.5.5 Detailed Audit Trails

To provide avenues for tracking and understanding of automated decisions retrospectively, protected and detailed logging mechanisms comparable to the black box of an aircraft are an option. Making such audit trails effective would require improving techniques to identify, prioritize, filter, summarize, retain, protect, organize, and make easily-consumable information about everything an agent encounters, infers, deduces, or decides [347].

6.2.5.6 Report Sufficiency

Quite a high volume of sensory and other data, computations involving processing intermediate values, and decisions of all sizes, will be present in any but the most trivial of agents. Specialized techniques are therefore required for determining the proper or sufficient amount of information to report to a human for them to be adequately informed but not overwhelmed. In the context of reporting on judgement, explaining the major factors that went into the decision and what would have to have been different for a different decision [219] might qualify, but the optimal level of detail will likely vary by application or context. The question arises as to the appropriate circumstances in which to produce a maximally informative report even when such an interpretable report necessarily impinges the accuracy or reliability of decisions made [93]. The agent would ideally determine the salience, risk, or uncertainty of a situation or a decision and modulate its introspection, logging, and reporting volume upward during times when those are elevated; furthermore, it could also modulate where, based on similar factors and operator expectations, on the pareto frontier between accuracy and interpretability to be [437]. When the report itself is interactive, such as via a dialogue, there is less uncertainty about sufficiency because humans can drill down for more detail where needed [339], so long as it has been logged in depth internally.

6.2.5.7 Causal Accounting

Automated reports that attempt to explain events often report on the correlations or covariances among different factors or features. Methods for disentangling correlation from causation, however, can clarify AI-generated explanations. When world models are too small or inadequately structured, it may not be possible. Work on generalizing the bounds in which one can isolate the effects of an action may be useful in wider use of such techniques [391]. See section *Causal Identification*, which addresses causal disentanglement for purposes of the agent's understanding, but which can also be leveraged for monitoring by operators.

7 Conclusion

In summary, success in the quest for artificial intelligence has the potential to bring unprecedented benefits to humanity, and it is therefore worthwhile to research how to maximize these benefits while avoiding potential pitfalls. This document has given numerous examples of such research, aimed at ensuring that AI remains robust and beneficial and aligned with human interests. Much of this research is early on and calls for further work. There are also likely to be many more topics and techniques needed for AI safety that have yet to be discovered.

References

[1] Pieter Abbeel, Adam Coates, and Andrew Y. Ng. "Autonomous Helicopter Aerobatics Through Apprenticeship Learning". In: *The International Journal of Robotics Research* 29.13 (2010), pp. 1608–1639. ISSN: 0278-3649. DOI: 10.1177/0278364910371999. URL: http://dx.doi.org/10.1177/0278364910371999.

[2] Pieter Abbeel and Andrew Y. Ng. "Apprenticeship Learning via Inverse Reinforcement Learning". In: *Proceedings of the Twenty-first International Conference on Machine Learning*. ICML '04. Banff, Alberta, Canada: ACM, 2004, pp. 1–. ISBN: 1-58113-838-5. DOI: 10.1145/1015330.1015430. URL: http://ai.stanford.edu/~ang/papers/icml04-apprentice.pdf.

[3] Pieter Abbeel and Andrew Y. Ng. "Exploration and Apprenticeship Learning in Reinforcement Learning". In: *Proceedings of the 22Nd International Conference on Machine Learning*. ICML '05. ACM, 2005, pp. 1–8. ISBN: 1-59593-180-5. DOI: 10.1145/1102351.1102352. URL: http://ai.stanford.edu/~pabbeel//pubs/AbbeelNg_eaalirl_ICML2005.pdf.

[4] Tudor Achim, Ashish Sabharwal, and Stefano Ermon. "Beyond Parity Constraints: Fourier Analysis of Hash Functions for Inference". In: *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*. Vol. 48. JMLR Workshop and Conference Proceedings. JMLR.org, 2016, pp. 2254–2262. URL: http://www.jmlr.org/proceedings/papers/v48/achim16.pdf.

[5] Alekh Agarwal et al. "Taming the monster: A fast and simple algorithm for contextual bandits". In: *In Proceedings of the 31st International Conference on Machine Learning (ICML-14*. 2014, pp. 1638–1646. URL: http://jmlr.org/proceedings/papers/v32/agarwalb14.pdf.

[6] Miltiadis Allamanis et al. "Learning Continuous Semantic Representations of Symbolic Expressions". In: *CoRR* abs/1611.01423 (2016). URL: http://arxiv.org/abs/1611.01423.

[7] Rajeev Alur. "Formal verification of hybrid systems". In: *Embedded Software (EMSOFT), 2011 Proceedings of the International Conference on*. IEEE. 2011, pp. 273–278. URL: https://www.cis.upenn.edu/~alur/EmsoftSurvey11.pdf.

[8] Rajeev Alur. *Principles of Cyber-Physical Systems*. MIT Press, 2015. ISBN: 9780262029117. URL: https://mitpress.mit.edu/books/principles-cyber-physical-systems.

[9] Saleema Amershi et al. "Power to the People: The Role of Humans in Interactive Machine Learning". In: *AI Magazine* (2014). URL: www.aaai.org/ojs/index.php/aimagazine/article/view/2513/2456.

[10] Kareem Amin and Satinder P. Singh. "Towards Resolving Unidentifiability in Inverse Reinforcement Learning". In: *CoRR* abs/1601.06569 (2016). URL: https://pdfs.semanticscholar.org/04cf/f669a73c4b6c84124de6e88562cab742c6cb.pdf.

[11] Dario Amodei et al. "Concrete Problems in AI Safety". In: *CoRR* abs/1606.06565 (2016). URL: http://arxiv.org/abs/1606.06565.

[12] Dario Amodei et al. "Deep Speech 2 : End-to-End Speech Recognition in English and Mandarin". In: *ICML*. Vol. 48. JMLR Workshop and Conference Proceedings. JMLR.org, 2016, pp. 173–182. URL: http://jmlr.org/proceedings/papers/v48/amodei16.pdf.

[13] Animashree Anandkumar, Daniel J. Hsu, and Sham M. Kakade. "A Method of Moments for Mixture Models and Hidden Markov Models". In: *CoRR* abs/1203.0683 (2012). URL: http://www.jmlr.org/proceedings/papers/v23/anandkumar12/anandkumar12.pdf.

[14] Michael Anderson and Susan Leigh Anderson, eds. *Machine Ethics*. 1st. Cambridge University Press, 2011. ISBN: 0521112354, 9780521112352.

[15] Michael Anderson and Susan Anderson. *GenEth: A General Ethical Dilemma Analyzer*. 2014. URL: http://www.aaai.org/ocs/index.php/AAAI/AAAI14/paper/download/8308/8428.

[16] Theodore W. Anderson and Herman Rubin. "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations". In: *The Annals of Mathematical Statistics* 20.1 (1949), pp. 46–63. DOI: 10.1214/aoms/1177730090. URL: http://dx.doi.org/10.1214/aoms/1177730090.

[17] Theodore W. Anderson and Herman Rubin. "The Asymptotic Properties of Estimates of the Parameters of a Single Equation in a Complete System of Stochastic Equations". In: *The Annals of Mathematical Statistics* 21.4 (1950), pp. 570–582. ISSN: 00034851. URL: http://www.jstor.org/stable/2236607.

[18] David Andre and Stuart J Russell. "State abstraction for programmable reinforcement learning agents". In: *Eighteenth national conference on Artificial intelligence*. American Association for Artificial Intelligence. 2002, pp. 119–125. URL: https://people.eecs.berkeley.edu/~russell/papers/aaai02-alisp.pdf.

[19] Stuart Armstrong. "Good and safe use of AI oracles". forthcoming paper. 2016. URL: https://dl.dropboxusercontent.com/u/23843264/Permanent/Using_Oracles.pdf.

[20] Stuart Armstrong. "Motivated Value Selection for Artificial Agents". In: *Artificial Intelligence and Ethics, Papers from the 2015 AAAI Workshop, Austin, Texas, USA, January 25, 2015.* 2015, pp. 12–20. URL: http://aaai.org/ocs/index.php/WS/AAAIW15/paper/viewFile/10183/10126.

[21] Stuart Armstrong. "Risks and Mitigation Strategies for Oracle AI". In: *Philosophy and Theory of Artificial Intelligence*. Ed. by Vincent C. Muller. Springer Berlin Heidelberg, 2013, pp. 335–347. ISBN: 978-3-642-31674-6. DOI: 10.1007/978-3-642-31674-6_25. URL: https://www.fhi.ox.ac.uk/wp-content/uploads/Risks-and-Mitigation-Strategies-for-Oracle-AI.pdf.

[22] Stuart Armstrong. *Utility Indifference*. Tech. rep. Future of Humanity Insitute, Oxford University, 2010. URL: http://www.fhi.ox.ac.uk/utility-indifference.pdf.

[23] Stuart Armstrong, Nick Bostrom, and Anders Sandberg. "Thinking Inside the Box: Controlling and Using an Oracle AI". In: *Minds and Machines* 22.4 (2012), pp. 299–324. URL: www.nickbostrom.com/papers/oracle.pdf.

[24] Stuart Armstrong and Jan Leike. *Towards Interactive Inverse Reinforcement Learning*. Presentation in (NIPS 2016) Workshop: Reliable Machine Learning In The Wild. 2016. URL: https://dl.dropboxusercontent.com/u/23843264/Permanent/towards-interactive-inverse-reinforcement-learning.pdf.

[25] Stuart Armstrong and Benjamin Levinstein. "Reduced Impact Artificial Intelligences". 2015. URL: https://dl.dropboxusercontent.com/u/23843264/Permanent/Reduced_impact_S+B.pdf.

[26] Peter M Asaro. "What should we want from a robot ethic?" In: *International Review of Information Ethics* 6.12 (2006), pp. 9–16. URL: http://www.peterasaro.org/writing/Asaro%20IRIE.pdf.

[27] Tamim Asfour et al. "Imitation learning of dual-arm manipulation tasks in humanoid robots". In: *International Journal of Humanoid Robotics* 5.02 (2008), pp. 183–202. URL: www.sfb588.uni-karlsruhe.de/Module/Publications/publications/Asfour2008a.pdf.

[28] Hutan Ashrafian. "Can Artificial Intelligences Suffer from Mental Illness? A Philosophical Matter to Consider". In: *Science and Engineering Ethics* (2016), pp. 1–10. ISSN: 1471-5546. DOI: 10.1007/s11948-016-9783-0. URL: http://link.springer.com/content/pdf/10.1007%2Fs11948-016-9783-0.pdf.

[29] Karl J Åström and Björn Wittenmark. *Adaptive control*. Courier Dover Publications, 2013.

[30] James Babcock, János Kramár, and Roman Yampolskiy. "The AGI Containment Problem". In: *Artificial General Intelligence: 9th International Conference, AGI 2016, New York, NY, USA, July 16-19, 2016, Proceedings*. Ed. by Bas Steunebrink, Pei Wang, and Ben Goertzel. Cham: Springer International Publishing, 2016, pp. 53–63. ISBN: 978-3-319-41649-6. DOI: 10.1007/978-3-319-41649-6_6. URL: https://pdfs.semanticscholar.org/d7a4/adb20a879e89fc12600c84dff0cb69fd7d58.pdf.

[31] Joscha Bach. *Discussion of generalized pathologies of intelligent agents and psychological and organizational psychological analogues*. Discussion. 2016.

[32] Joscha Bach. *Modeling Emotion as an Interaction between Motivation and Modulated Cognition.* 2011. URL: http://www.lorentzcenter.nl/lc/web/2011/464/presentations/Bach.pdf.

[33] David Baehrens et al. "How to explain individual classification decisions". In: *The Journal of Machine Learning Research* 11 (2010), pp. 1803–1831. URL: http://is.tuebingen.mpg.de/fileadmin/user_upload/files/publications/baehrens10a_%5B0%5D.pdf.

[34] Aijun Bai, Siddharth Srivastava, and Stuart J. Russell. "Markovian State and Action Abstractions for MDPs via Hierarchical Monte Carlo Tree Search". In: *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*. 2016, pp. 3029–3039. URL: http://www.ijcai.org/Abstract/16/430.

[35] Christel Baier and Joost-Pieter Katoen. *Principles of Model Checking (Representation and Mind Series)*. The MIT Press, 2008. ISBN: 026202649X, 9780262026499.

[36] Chris L. Baker, Rebecca R. Saxe, and Joshua B. Tenenbaum. "Bayesian theory of mind: Modeling joint belief-desire attribution". In: *CogSci 2011 Proceedings*. 2011. URL: http://aiweb.cs.washington.edu/research/projects/aiweb/media/papers/cogsci2011.pdf.

[37] Chris L. Baker, Rebecca Saxe, and Joshua B. Tenenbaum. "Action understanding as inverse planning". In: *Cognition* 113.3 (2009). Reinforcement learning and higher cognition, pp. 329–349. ISSN: 0010-0277. DOI: http://dx.doi.org/10.1016/j.cognition.2009.07.005. URL: http://web.mit.edu/clbaker/www/papers/cognition2009.pdf.

[38] Krishnakumar Balasubramanian, Pinar Donmez, and Guy Lebanon. "Unsupervised Supervised Learning II: Margin-Based Classification Without Labels". In: *J. Mach. Learn. Res.* 12 (Nov. 2011), pp. 3119–3145. ISSN: 1532-4435. URL: http://www.jmlr.org/papers/volume12/balasubramanian11a/balasubramanian11a.pdf.

[39] Mihály Bárász et al. "Robust Cooperation in the Prisoner's Dilemma: Program Equilibrium via Provability Logic". In: *CoRR* abs/1401.5577 (2014). URL: https://arxiv.org/pdf/1401.5577v1.

[40] Marco Barreno et al. "The Security of Machine Learning". In: *Machine Learning* 81.2 (Nov. 2010), pp. 121–148. ISSN: 0885-6125. DOI: 10.1007/s10994-010-5188-5. URL: http://dx.doi.org/10.1007/s10994-010-5188-5.

[41] Anthony Barrett and Seth Baum. "A Model of Pathways to Artificial Superintelligence Catastrophe for Risk and Decision Analysis". In: *Journal of Experimental Theoretical Artificial Intelligence* (2016). URL: http://sethbaum.com/ac/fc_AI-Pathways.pdf.

[42] Sridevi Baskaran et al. "Efficient Discovery of Ontology Functional Dependencies". In: *CoRR* abs/1611.02737 (2016). URL: https://arxiv.org/pdf/1611.02737v2.

[43] C. Beattie et al. "DeepMind Lab". In: *ArXiv e-prints* (Dec. 2016). arXiv: 1612.03801 [cs.AI]. URL: https://arxiv.org/pdf/1612.03801v2.

[44] N. Ben Zrihem, T. Zahavy, and S. Mannor. "Visualizing Dynamics: from t-SNE to SEMI-MDPs". In: *ArXiv e-prints* (June 2016). arXiv: 1606.07112 [stat.ML]. URL: https://icmlviz.github.io/assets/papers/10.pdf.

[45] Tsvi Benson-Tilsen. *Training Garrabrant inductors to predict counterfactuals*. Blog post. 2016. URL: https://agentfoundations.org/item?id=1054.

[46] Tsvi Benson-Tilsen. *UDT with known search order*. Tech. rep. Machine Intelligence Research Institute, 2014. URL: http://intelligence.org/files/UDTSearchOrder.pdf.

[47] Tsvi Benson-Tilsen and Nate Soares. "Formalizing Convergent Instrumental Goals". In: *The Workshops of the Thirtieth AAAI Conference on Artificial Intelligence*. 2016, pp. 62–70. URL: https://intelligence.org/files/FormalizingConvergentGoals.pdf.

[48] Felix Berkenkamp, Andreas Krause, and Angela P. Schoellig. *Bayesian optimization with safety constraints: safe and automatic parameter tuning in robotics*. Tech. rep. 2016. URL: https://arxiv.org/pdf/1602.04450.pdf.

[49] Francesco Berto. "Impossible Worlds". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2013. Metaphysics Research Lab, Stanford University, 2013. URL: https://plato.stanford.edu/archives/win2013/entries/impossible-worlds/.

[50] Dimitris Bertsimas, David B. Brown, and Constantine Caramanis. "Theory and Applications of Robust Optimization". In: *SIAM Rev.* 53.3 (Aug. 2011), pp. 464–501. ISSN: 0036-1445. DOI: 10.1137/080734510. URL: https://faculty.fuqua.duke.edu/~dbbrown/bio/papers/bertsimas_brown_caramanis_11.pdf.

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

[52] Alina Beygelzimer et al. "Search Improves Label for Active Learning". In: *Advances In Neural Information Processing Systems 29*. Ed. by D. D. Lee et al. Curran Associates, Inc., 2016, pp. 3342–3350. URL: http://papers.nips.cc/paper/6182-search-improves-label-for-active-learning.pdf.

[53] Or Biran and Kathleen McKeown. "Justification Narratives for Individual classifications". In: *Proceedings of the AutoML Workshop at ICML 2014 (2014)*. 2014. URL: http://www.cs.columbia.edu/~orb/papers/justification_automl_2014.pdf.

[54] Jon Bird and Paul Layzell. "The Evolved Radio and Its Implications for Modelling the Evolution of Novel Sensors". In: *Proceedings of the Evolutionary Computation on 2002. CEC '02. Proceedings of the 2002 Congress - Volume 02*. CEC '02. IEEE Computer Society, 2002, pp. 1836–1841. ISBN: 0-7803-7282-4. DOI: 10.1109/CEC.2002.1004522. URL: https://people.duke.edu/~ng46/topics/evolved-radio.pdf.

[55] Peter de Blanc. "Ontological Crises in Artificial Agents' Value Systems". In: *CoRR* abs/1105.3821 (2011). URL: http://arxiv.org/abs/1105.3821.

[56] Charles Blundell et al. "Weight Uncertainty in Neural Networks". In: *Proceedings of The 32nd International Conference on Machine Learning*. JMLR, 2015, pp. 1613–1622. URL: http://jmlr.org/proceedings/papers/v37/blundell15.pdf.

[57] Zahy Bnaya et al. "Confidence Backup Updates for Aggregating MDP State Values in Monte-Carlo Tree Search". In: *Eighth Annual Symposium on Combinatorial Search*. 2015. URL: www.aaai.org/ocs/index.php/SOCS/SOCS15/paper/view/11273/10648.

[58] Nick Bostrom. *Superintelligence: Paths, dangers, strategies*. Oxford University Press, 2014.

[59] Nick Bostrom. "The superintelligent will: Motivation and instrumental rationality in advanced artificial agents". In: *Minds and Machines* 22.2 (2012), pp. 71–85. URL: http://www.nickbostrom.com/superintelligentwill.pdf.

[60] Nick Bostrom et al. "Infinite ethics". In: *Analysis and Metaphysics* 10 (2011), pp. 9–59. URL: http://www.nickbostrom.com/ethics/infinite.pdf.

[61] Leon Bottou. "Two big challenges in machine learning". In: Invited talk at ICML 32nd International Conference on Machine Learning 2015 - LILLE. 2015. URL: http://icml.cc/2015/invited/LeonBottouICML2015.pdf.

[62] Léon Bottou et al. *Counterfactual Reasoning and Learning Systems*. Tech. rep. arXiv:1209.2355, 2012. URL: http://leon.bottou.org/papers/tr-bottou-2012.

[63] Craig Boutilier. "A POMDP Formulation of Preference Elicitation Problems". In: *Eighteenth National Conference on Artificial Intelligence*. Edmonton, Alberta, Canada: American Association for Artificial Intelligence, 2002, pp. 239–246. ISBN: 0-262-51129-0. URL: https://www.aaai.org/Papers/AAAI/2002/AAAI02-037.pdf.

[64] Daniel J. Brooks et al. "Towards State Summarization for Autonomous Robots". In: *Dialog with Robots, Papers from the 2010 AAAI Fall Symposium, Arlington, Virginia, USA, November 11-13, 2010*. AAAI, 2010. URL: http://www.aaai.org/ocs/index.php/FSS/FSS10/paper/viewFile/2223/2749.

[65] Achim D Brucker and Uwe Sodan. "Deploying static application security testing on a large scale". In: *GI Sicherheit 2014*. Vol. 228. GI. 2014, pp. 91–101. URL: https://www.brucker.ch/bibliography/download/2014/brucker.ea-sast-expierences-2014.pdf.

[66] Yuriy Brun and Michael D Ernst. "Finding latent code errors via machine learning over program executions". In: *Proceedings of the 26th International Conference on Software Engineering*. IEEE Computer Society. 2004, pp. 480–490. URL: https://homes.cs.washington.edu/~mernst/pubs/machlearn-errors-icse2004.pdf.

[67] Miles Brundage. "Limitations and risks of machine ethics". In: *Journal of Experimental Theoretical Artificial Intelligence* 26.3 (2014), pp. 355–372. DOI: 10.1080/0952813X.2014.895108. eprint: http://dx.doi.org/10.1080/0952813X.2014.895108. URL: http://www.milesbrundage.com/uploads/2/1/6/8/21681226/limitations_and_risks_of_machine_ethics.pdf.

[68] Lucian Busoniu, Robert Babuvska, and Bart De Schutter. "A Comprehensive Survey of Multiagent Reinforcement Learning". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 38.2 (2008), pp. 156–172. ISSN: 1094-6977. DOI: 10.1109/TSMCC.2007.913919. URL: http://repository.tudelft.nl/assets/uuid:4c7d3b49-06fc-400c-923e-3903b8d230fe/busoniu2008.pdf.

[69] Tamara Carleton, William Cockayne, and Larry Leifer. *An Exploratory Study about the Role of Ambiguity During Complex Problem Solving.* 2007. URL: https://www.aaai.org/Papers/Symposia/Spring/2008/SS-08-03/SS08-03-002.pdf.

[70] Giovanni Sirio Carmantini et al. "A modular architecture for transparent computation in Recurrent Neural Networks". In: *CoRR* abs/1609.01926 (2016). URL: http://arxiv.org/abs/1609.01926.

[71] José Carmo and Andrew J. I. Jones. "Deontic database constraints, violation and recovery". In: *Studia Logica* 57.1 (1996), pp. 139–165. ISSN: 1572-8730. DOI: 10.1007/BF00370673. URL: http://dx.doi.org/10.1007/BF00370673.

[72] Ivan Sanchez Carmona and Sebastian Riedel. "Extracting Interpretable Models from Matrix Factorization Models". In: *Proceedings of the 2015th International Conference on Cognitive Computation: Integrating Neural and Symbolic Approaches - Volume 1583.* COCO'15. CEUR-WS.org, 2015, pp. 78–84. URL: http://ceur-ws.org/Vol-1583/CoCoNIPS_2015_paper_10.pdf.

[73] David Chalmers. "The singularity: A philosophical analysis". In: *Journal of Consciousness Studies* 17.9-10 (2010), pp. 7–65. URL: http://consc.net/papers/singularity.pdf.

[74] M. B. Chang et al. "A Compositional Object-Based Approach to Learning Physical Dynamics". In: *ArXiv e-prints* (Dec. 2016). arXiv: 1612.00341 [cs.AI]. URL: http://phys.csail.mit.edu/papers/11.pdf.

[75] Angela Chen. *More Rational Resolutions.* Blog Post. 2014. URL: http://www.wsj.com/articles/SB10001424052702303453004579290510733740616.

[76] Li Chen and Pearl Pu. *Survey of Preference Elicitation Methods.* Tech. rep. 2004. URL: https://infoscience.epfl.ch/record/52659/files/IC_TECH_REPORT_200467.pdf.

[77] Xiangli Chen et al. "Robust Covariate Shift Regression." In: *AISTATS.* Ed. by Arthur Gretton and Christian C. Robert. Vol. 51. JMLR Workshop and Conference Proceedings. JMLR.org, 2016, pp. 1270–1279. URL: http://dblp.uni-trier.de/db/conf/aistats/aistats2016.html#ChenMLZ16.

[78] Paul Christiano. *ALBA: An explicit proposal for aligned AI.* 2016. URL: https://medium.com/ai-control/alba-an-explicit-proposal-for-aligned-ai-17a55f60bbcf.

[79] Paul Christiano. *Active learning for opaque, powerful predictors.* Blog Post. 2015. URL: https://medium.com/ai-control/active-learning-for-opaque-powerful-predictors-94724b3adf06.

[80] Paul Christiano. *Ambitious vs. narrow value learning.* Medium Corporation, 2015.

[81] Paul Christiano. *Ambitious vs. narrow value learning.* Blog Post. 2015. URL: https://medium.com/ai-control/ambitious-vs-narrow-value-learning-99bd0c59847e.

[82] Paul Christiano. *Approval-directed Agents.* Blog Post. 2014. URL: https://medium.com/ai-control/model-free-decisions-6e6609f5d99e.

[83] Paul Christiano. *Approval-directed algorithm learning.* Blog Post. 2015. URL: https://medium.com/ai-control/approval-directed-algorithm-learning-bf1f8fad42cd.

[84] Paul Christiano. *Capability amplification.* Medium Corporation, 2016.

[85] Paul Christiano. *Learning with catastrophes.* Medium Corporation, 2016. URL: https://medium.com/ai-control/learning-with-catastrophes-59387b55cc30.

[86] Paul Christiano. *Mimicry and meeting halfway.* Blog Post. 2015. URL: https://medium.com/ai-control/mimicry-maximization-and-meeting-halfway-c149dd23fc17.

[87] Paul Christiano. *Non-Omniscience, Probabilistic Inference, and Metamathematics*. Tech. rep. Machine Intelligence Research Institute, 2014. URL: intelligence.org/files/Non-Omniscience.pdf.

[88] Paul Christiano. *Reward engineering*. Medium Corporation, 2015. URL: https://medium.com/ai-control/reward-engineering-f8b5de40d075.

[89] Paul Christiano. *Scalable AI control*. Blog Post. 2015. URL: https://medium.com/ai-control/scalable-ai-control-7db2436feee7.

[90] Paul Christiano. *Semi-supervised reinforcement learning*. Blog Post. 2016. URL: https://medium.com/ai-control/semi-supervised-reinforcement-learning-cf7d5375197f.

[91] Paul Christiano. *Specifying 'Enlightened Judgment' Precisely (Reprise)*. Blog Post. 2014. URL: https://ordinaryideas.wordpress.com/2014/08/27/specifying-enlightened-judgment-precisely-reprise/.

[92] Paul Christiano. *The informed oversight problem*. Blog Post. 2016. URL: https://medium.com/ai-control/the-informed-oversight-problem-1b51b4f66b35.

[93] Paul Christiano. *The informed oversight problem*. Medium Corporation, 2016. URL: https://medium.com/ai-control/the-informed-oversight-problem-1b51b4f66b35.

[94] Wei Chu and Zoubin Ghahramani. "Preference learning with Gaussian processes". In: *Proceedings of the 22nd international conference on Machine learning*. ACM. 2005, pp. 137–144. URL: http://www.gatsby.ucl.ac.uk/%7Echuwei/paper/gppl.pdf.

[95] Edmund Clarke, Tom Henzinger, and Veith Helmut. "Handbook of Model Checking". In: (to appear). Springer, 2017.

[96] Roberta Coelho et al. "Unit testing in multi-agent systems using mock agents and aspects". In: *Proceedings of the 2006 international workshop on Software engineering for large-scale multi-agent systems*. ACM. 2006, pp. 83–90. URL: https://pdfs.semanticscholar.org/363c/c023e00467141712292d9ecafa15acd78b25.pdf.

[97] Paul R. Cohen et al. "Robot Baby 2001". In: *Algorithmic Learning Theory: 12th International Conference, ALT 2001 Washington, DC, USA, November 25–28, 2001 Proceedings*. Ed. by Naoki Abe, Roni Khardon, and Thomas Zeugmann. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001, pp. 32–56. ISBN: 978-3-540-45583-7. DOI: 10.1007/3-540-45583-3_4. URL: http://www-symbiotic.cs.ou.edu/~fagg/classes/neurocog/restrict/papers/cohen_etal_2001.pdf.

[98] Kenneth Mark Colby. *Artificial paranoia: A computer simulation of paranoid processes*. Vol. 49. Elsevier, 2013. URL: http://www.csee.umbc.edu/courses/graduate/CMSC671/fall12/resources/colby_71.pdf.

[99] Yann Collette and Patrick Slarry. *Multiobjective Optimization*. Decision Engineering. Springer, 2004. DOI: 10.1007/978-3-662-08883-8.

[100] N. Condry. "Meaningful Models: Utilizing Conceptual Structure to Improve Machine Learning Interpretability". In: *ArXiv e-prints* (2016). arXiv: 1607.00279 [stat.ML]. URL: https://arxiv.org/pdf/1607.00279.pdf.

[101] Peter Congdon. *Applied Bayesian Modeling*. 2nd. Wiley, 2014.

[102] Vince Conitzer, Markus Brill, and Rupert Freeman. "Crowdsourcing Societal Tradeoffs". In: *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*. AAMAS '15. Istanbul, Turkey: International Foundation for Autonomous Agents and Multiagent Systems, 2015, pp. 1213–1217. ISBN: 978-1-4503-3413-6. URL: https://www.cs.duke.edu/~rupert/crowdsourcing.pdf.

[103] Vincent Conitzer. "Metareasoning as a Formal Computational Problem". In: *Metareasoning: Thinking About Thinking*. Ed. by Michael Cox and Anita Raja. MIT Press, 2011, pp. 119–128. ISBN: 9780262295284.

[104] Vincent Conitzer and Tuomas Sandholm. "Definition and Complexity of Some Basic Metareasoning Problems". In: *Proceedings of the 18th International Joint Conference on Artificial Intelligence*. IJCAI'03. Acapulco, Mexico: Morgan Kaufmann Publishers Inc., 2003, pp. 1099–1106. URL: https://www.cs.cmu.edu/~sandholm/complexity_of_metareasoning.ijcai03.pdf.

[105] Vincent Conitzer et al. "Moral Decision Making Frameworks for Artificial Intelligence". In: (2017). To appear in Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17) Senior Member / Blue Sky Track, San Francisco, CA, USA, 2017. URL: https://users.cs.duke.edu/~conitzer/moralAAAI17.pdf.

[106] Vincent Conitzer et al. *Rules for Choosing Societal Tradeoffs*. 2016. URL: http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12504/11620.

[107] Ariel Conn. *Training Artificial Intelligence to Compromise*. Blog Post. 2016. URL: http://futureoflife.org/2016/09/26/training-artificial-intelligence-compromise/.

[108] Mark G. Core et al. "Building explainable artificial intelligence systems". In: *In Proceedings of the 18th Conference on Innovative Applications of Artificial Intelligence (IAAI-06)*. 2006. URL: http://www.aaai.org/Papers/IAAI/2006/IAAI06-010.pdf.

[109] Michael Cox and Anita Raja. *Metareasoning : thinking about thinking*. MIT Press, 2011, p. 352. ISBN: 9780262295284.

[110] Fabio Cozman and Ira Cohen. "Risks of semi-supervised learning". In: *Semi-supervised learning* (2006), pp. 56–72.

[111] Luis G. Crespo, Megumi Matsutani, and Anuradha M. Annaswamy. *Verification and Tuning of an Adaptive Controller for an Unmanned Air Vehicle*. 2010. URL: http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20100031105.pdf.

[112] Andrew Critch. *Discussion of one algorithm controlling another and open source game theory*. Discussion. 2016.

[113] Andrew Critch. "Parametric Bounded Löb's Theorem and Robust Cooperation of Bounded Agents". In: *CoRR* abs/1602.04184 (2016). URL: http://arxiv.org/abs/1602.04184.

[114] Fiery Cushman. "Action, Outcome, and Value: A Dual System Framework for Morality". In: *Personality and social psychology review* 17.3 (2013), pp. 273–292. URL: http://cushmanlab.fas.harvard.edu/docs/cushman_2013.pdf.

[115] DARPA. "DARPA Seeks to Remove Communication Barrier Between Humans and Computers". In: (2015). URL: http://www.darpa.mil/news-events/2015-02-20.

[116] V. D'Silva, D. Kroening, and G. Weissenbacher. "A Survey of Automated Techniques for Formal Software Verification". In: *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 27.7 (2008), pp. 1165–1178. ISSN: 0278-0070. DOI: 10.1109/TCAD.2008.923410. URL: http://www.kroening.com/papers/tcad-sw-2008.pdf.

[117] Christian Daniel et al. "Active reward learning with a novel acquisition function". In: *Autonomous Robots* 39.3 (2015), pp. 389–405. ISSN: 1573-7527. DOI: 10.1007/s10514-015-9454-z. URL: http://www.ausy.tu-darmstadt.de/uploads/Team/ChristianDaniel/ActiveRewardLearning.pdf.

[118] Rajarshi Das et al. "Chains of Reasoning over Entities, Relations, and Text using Recurrent Neural Networks". In: *CoRR* abs/1607.01426 (2016). URL: http://arxiv.org/abs/1607.01426.

[119] Anupam Datta, Shayak Sen, and Yair Zick. "Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems". In: *2016 IEEE Symposium on Security and Privacy (SP)*. 2016, pp. 598–617. DOI: 10.1109/SP.2016.42. URL: https://www.andrew.cmu.edu/user/danupam/datta-sen-zick-oakland16.pdf.

[120] Ernest Davis. "Ethical guidelines for a superintelligence". In: *Artificial Intelligence* 220 (2015), pp. 121–124. DOI: 10.1016/j.artint.2014.12.003. URL: http://dx.doi.org/10.1016/j.artint.2014.12.003.

[121] Alexander Philip Dawid and Allan M Skene. "Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm". In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28.1 (1979), pp. 20–28. ISSN: 00359254, 14679876. URL: http://www.jstor.org/stable/2346806.

[122] Peter Dayan and Geoffrey E. Hinton. "Feudal Reinforcement Learning". In: *Advances in Neural Information Processing Systems 5, [NIPS Conference]*. Morgan Kaufmann Publishers Inc., 1993, pp. 271–278. ISBN: 1-55860-274-7. URL: https://papers.nips.cc/paper/714-feudal-reinforcement-learning.pdf.

[123] André DeHon et al. "Preliminary design of the SAFE platform". In: *Proceedings of the 6th Workshop on Programming Languages and Operating Systems*. ACM. 2011, p. 4. URL: http://repository.upenn.edu/cgi/viewcontent.cgi?article=1707&context=cis_papers.

[124] Kalyanmoy Deb. "Multi-objective Optimization". In: *Search Methodologies: Introductory Tutorials in Optimization and Decision Support Techniques*. Ed. by Edmund K. Burke and Graham Kendall. Springer US, 2014, pp. 403–449. ISBN: 978-1-4614-6940-7. DOI: 10.1007/978-1-4614-6940-7_15. URL: http://dx.doi.org/10.1007/978-1-4614-6940-7_15.

[125] Morteza Dehghani et al. "An Integrated Reasoning Approach to Moral Decision-Making". In: *Machine Ethics*. Ed. by Michael Anderson and Susan Leigh Anderson. Cambridge University Press, May 2011. URL: http://ict.usc.edu/pubs/Machine%20Ethics.pdf.

[126] Abram Demski. "Logical Prior Probability". In: *Artificial General Intelligence: 5th International Conference, AGI 2012, Oxford, UK, December 8-11, 2012. Proceedings*. Ed. by Joscha Bach, Ben Goertzel, and Matthew Iklé. Springer Berlin Heidelberg, 2012, pp. 50–59. ISBN: 978-3-642-35506-6. DOI: 10.1007/978-3-642-35506-6_6. URL: http://ict.usc.edu/pubs/Logical%20Prior%20Probability.pdf.

[127] Louise A Dennis et al. "Practical Verification of Decision-Making in Agent-Based Autonomous Systems". In: *arXiv preprint arXiv:1310.2431* (2013). URL: http://repository.liv.ac.uk/13195/1/verification_arxiv.pdf.

[128] Daniel Dewey. "Learning What to Value". In: *Artificial General Intelligence: 4th International Conference, AGI 2011, Mountain View, CA, USA, August 3-6, 2011. Proceedings*. Ed. by Jurgen Schmidhuber, Kristinn R. Thorisson, and Moshe Looks. Vol. 6830. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 309–314. ISBN: 978-3-642-22887-2. DOI: 10.1007/978-3-642-22887-2_35. URL: http://www.danieldewey.net/learning-what-to-value.pdf.

[129] Daniel Dewey. "Reinforcement Learning and the Reward Engineering Principle". In: *2014 AAAI Spring Symposium Series*. 2014. URL: http://www.danieldewey.net/reward-engineering-principle.pdf.

[130] Tom Dietterich and Eric Horvitz. *Benefits and Risks of Artificial Intelligence*. blog post. 2015.

[131] United Nations Institute for Disarmament Research. *The Weaponization of Increasingly Autonomous Technologies: Implications for Security and Arms Control*. UNIDIR, 2014.

[132] DoD. *Department of Defense Trusted Computer System Evaluation Criteria*. Department of Defense Standard DOD 5200.28- STD. 1985. URL: http://csrc.nist.gov/publications/history/dod85.pdf.

[133] József Dombi and Nándor J. Vincze. "Universal characterization of non-transitive preferences". In: *Mathematical Social Sciences* 27.1 (1994), pp. 91 –104. ISSN: 0165-4896. DOI: http://dx.doi.org/10.1016/0165-4896(94)00735-7. URL: http://www.sciencedirect.com/science/article/pii/0165489694007357.

[134] Pinar Donmez, Guy Lebanon, and Krishnakumar Balasubramanian. "Unsupervised Supervised Learning I: Estimating Classification and Regression Errors Without Labels". In: *The Journal of Machine Learning Research* 11 (Aug. 2010), pp. 1323–1351. ISSN: 1532-4435. URL: http://dl.acm.org/citation.cfm?id=1756006.1859895.

[135] Gregory Druck, Gideon Mann, and Andrew McCallum. "Learning from Labeled Features Using Generalized Expectation Criteria". In: *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '08. ACM, 2008, pp. 595–602. ISBN: 978-1-60558-164-4. DOI: 10.1145/1390334.1390436. URL: http://doi.acm.org/10.1145/1390334.1390436.

[136] Owain Evans, Andreas Stuhlmueller, and Noah D. Goodman. "Learning the Preferences of Bounded Agents". In: (2015). URL: http://stuhlmueller.org/papers/preferences-nipsworkshop2015.pdf.

[137] Owain Evans, Andreas Stuhlmüller, and Noah D. Goodman. "Learning the Preferences of Ignorant, Inconsistent Agents". In: *CoRR* abs/1512.05832 (2015). URL: http://arxiv.org/abs/1512.05832.

[138] Tom Everitt, and Marcus Hutter. “Avoiding Wireheading with Value Reinforcement Learning”. In: *Artificial General Intelligence: 9th International Conference, AGI 2016, New York, NY, USA, July 16-19, 2016, Proceedings*. Ed. by Bas Steunebrink, Pei Wang, and Ben Goertzel. Springer International Publishing, 2016, pp. 12–22. ISBN: 978-3-319-41649-6. DOI: 10.1007/978-3-319-41649-6_2. URL: http://www.springer.com/cda/content/document/cda_downloaddocument/9783319416489-c2.pdf.

[139] Tom Everitt et al. “Self-Modification of Policy and Utility Function in Rational Agents”. In: *Artificial General Intelligence: 9th International Conference, AGI 2016, New York, NY, USA, July 16-19, 2016, Proceedings*. Ed. by Bas Steunebrink, Pei Wang, and Ben Goertzel. Springer International Publishing, 2016, pp. 1–11. ISBN: 978-3-319-41649-6. DOI: 10.1007/978-3-319-41649-6_1. URL: http://www.tomeveritt.se/papers/AGI16-sm.pdf.

[140] Benja Fallenstein and Ramana Kumar. *Proof-producing reflection for HOL with an application to model polymorphism*. Springer, 2015. URL: https://www.cl.cam.ac.uk/~rk436/itp2015a.pdf.

[141] Benja Fallenstein and Nate Soares. “Problems of Self-reference in Self-improving Space-Time Embedded Intelligence”. In: *Artificial General Intelligence*. Lecture Notes in Computer Science. Springer International Publishing, 2014, pp. 21–32. URL: https://intelligence.org/files/ProblemsSelfReference.pdf.

[142] Benja Fallenstein and Nate Soares. *Vingean Reflection: Reliable Reasoning for Self-Modifying Agents*. Tech. rep. Machine Intelligence Research Institute, 2015. URL: https://intelligence.org/files/VingeanReflection.pdf.

[143] Benja Fallenstein, Nate Soares, and Jessica Taylor. “Reflective Variants of Solomonoff Induction and AIXI”. In: *In Artificial General Intelligence: 8th International Conference, AGI 2015*. Machine Intelligence Research Institute. Springer International Publishing, 2015, pp. 60–69. URL: https://intelligence.org/files/ReflectiveSolomonoffAIXI.pdf.

[144] Benja Fallenstein and Nisan Stiennon. *”Loudness: On priors over preference relations (Brief technical note)”*. Tech. rep. 2014. URL: https://intelligence.org/files/LoudnessPriors.pdf.

[145] Benja Fallenstein, Jessica Taylor, and Paul F. Christiano. “Reflective Oracles: A Foundation for Game Theory in Artificial Intelligence”. In: *Logic, Rationality, and Interaction: 5th International Workshop, LORI 2015, Taipei, Taiwan, October 28-30, 2015. Proceedings*. Ed. by Wiebe van der Hoek, Wesley H. Holliday, and Wen-fang Wang. Springer Berlin Heidelberg, 2015, pp. 411–415. ISBN: 978-3-662-48561-3. DOI: 10.1007/978-3-662-48561-3_34. URL: https://intelligence.org/files/ReflectiveOraclesAI.pdf.

[146] Amir M. Farahmand et al. “Regularized Policy Iteration”. In: *Advances in Neural Information Processing Systems 21*. Ed. by D. Koller et al. Curran Associates, Inc., 2009, pp. 441–448. URL: http://papers.nips.cc/paper/3445-regularized-policy-iteration.pdf.

[147] Rizal Fathony et al. “Adversarial Multiclass Classification: A Risk Minimization Perspective”. In: *Advances in Neural Information Processing Systems 29*. Ed. by D. D. Lee et al. Curran Associates, Inc., 2016, pp. 559–567. URL: http://papers.nips.cc/paper/6088-adversarial-multiclass-classification-a-risk-minimization-perspective.pdf.

[148] C. Finn et al. “Generalizing Skills with Semi-Supervised Reinforcement Learning”. In: *ArXiv e-prints* (Dec. 2016). eprint: 1612.00429. URL: https://arxiv.org/pdf/1612.00429v1.

[149] Chelsea Finn, Sergey Levine, and Pieter Abbeel. “Guided Cost Learning: Deep Inverse Optimal Control via Policy Optimization”. In: *Proceedings of the 33rd International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: WCP volume 48*. JMLR.org, 2016, pp. 49–58. URL: http://jmlr.org/proceedings/papers/v48/finn16.pdf.

[150] Kathleen Fisher. “HACMS: high assurance cyber military systems”. In: *Proceedings of the 2012 ACM conference on high integrity language technology*. ACM. 2012, pp. 51–52.

[151] Cameron E. Freer, Daniel M. Roy, and Joshua B. Tenenbaum. “Towards common-sense reasoning via conditional simulation: legacies of Turing in Artificial Intelligence”. In: *CoRR* abs/1212.4799 (2012). URL: https://arxiv.org/pdf/1212.4799v2.

[152] Robert M. French. "Catastrophic Interference in Connectionist Networks: Can It Be Predicted, Can It Be Prevented?" In: *Proceedings of the 6th International Conference on Neural Information Processing Systems*. NIPS'93. Denver, Colorado: Morgan Kaufmann Publishers Inc., 1993, pp. 1176–1177. URL: https://papers.nips.cc/paper/799-catastrophic-interference-in-connectionist-networks-can-it-be-predicted-can-it-be-prevented.pdf.

[153] Nir Friedman, Dan Geiger, and Moises Goldszmidt. "Bayesian Network Classifiers". In: *Machine Learning* 29.2-3 (Nov. 1997), pp. 131–163. ISSN: 0885-6125. DOI: 10.1023/A:1007465528199. URL: http://dx.doi.org/10.1023/A:1007465528199.

[154] Nathan Fulton and André Platzer. "A Logic of Proofs for Differential Dynamic Logic: Toward Independently Checkable Proof Certificates for Dynamic Logics". In: *Proceedings of the 2016 Conference on Certified Programs and Proofs, CPP 2016, St. Petersburg, FL, USA, January 18-19, 2016*. Ed. by Jeremy Avigad and Adam Chlipala. ACM, 2016, pp. 110–121. DOI: 10.1145/2854065.2854078. URL: http://nfulton.org/papers/lpdl.pdf.

[155] Haim Gaifman. "Reasoning with Limited Resources and Assigning Probabilities to Arithmetical Statements". In: *Synthese* 140.1 (2004), pp. 97–119. ISSN: 1573-0964. DOI: 10.1023/B:SYNT.0000029944.99888.a7. URL: http://www.columbia.edu/~hg17/synthese-paper-as-published.pdf.

[156] Yarin Gal and Zoubin Ghahramani. "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning". In: *Proceedings of the 33rd International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: WCP volume 48*. JMLR.org, 2016, pp. 1050–1059. URL: http://jmlr.org/proceedings/papers/v48/gal16-supp.pdf.

[157] C. Gallego-Ortiz and A. L. Martel. "Interpreting extracted rules from ensemble of trees: Application to computer-aided diagnosis of breast MRI". In: *ArXiv e-prints* (June 2016). arXiv: 1606.08288 [stat.ML]. URL: https://pdfs.semanticscholar.org/748d/3ef9f05b3c95cf82a14aa64549bcbe94be60.pdf.

[158] João Gama et al. "Learning with Drift Detection." In: *SBIA*. Ed. by Ana L. C. Bazzan and Sofiane Labidi. Vol. 3171. Lecture Notes in Computer Science. Springer, 2004, pp. 286–295. ISBN: 3-540-23237-0. URL: http://dx.doi.org/10.1007/978-3-540-28645-5_29.

[159] Yaroslav Ganin et al. "Domain-adversarial Training of Neural Networks". In: *The Journal of Machine Learning Research* 17.1 (Jan. 2016), pp. 2096–2030. ISSN: 1532-4435. URL: http://jmlr.org/papers/volume17/15-239/15-239.pdf.

[160] Javier Garcia and Fernando Fernandez. "A comprehensive survey on safe reinforcement learning". In: *The Journal of Machine Learning Research* (2015), pp. 1437–1480. URL: http://www.jmlr.org/papers/volume16/garcia15a/garcia15a.pdf.

[161] Peter Gardenfors. *Conceptual Spaces: The Geometry of Thought*. MIT Press, 2000, p. 317. ISBN: 9780262071994.

[162] Marta Garnelo, Kai Arulkumaran, and Murray Shanahan. "Towards Deep Symbolic Reinforcement Learning". In: *CoRR* abs/1609.05518 (2016). URL: http://arxiv.org/abs/1609.05518.

[163] Scott Garrabrant et al. *Logical Induction (Abridged)*. Tech. rep. 2016. URL: https://intelligence.org/files/LogicalInductionAbridged.pdf.

[164] Scott Garrabrant et al. "Logical Induction". In: *CoRR* abs/1609.03543 (2016). URL: http://arxiv.org/abs/1609.03543.

[165] Rayid Ghani. *You Say You Want Transparency and Interpretability?* 2016. URL: http://www.rayidghani.com/you-say-you-want-transparency-and-interpretability.

[166] Shalini Ghosh et al. "Trusted Machine Learning for Probabilistic Models". In: (2016). Contributed Talk: Reliable Machine Learning in the Wild at ICML 2016. URL: https://4caf2f9f-a-62cb3a1a-s-sites.googlegroups.com/site/wildml2016/ghosh16trusted.pdf.

[167] Alec Go, Richa Bhayani, and Lei Huang. "Twitter sentiment classification using distant supervision". In: *CS224N Project Report, Stanford* 1 (2009), p. 12. URL: https://www-cs.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf.

[168] B. Goertzel et al. *Real-World Reasoning: Toward Scalable, Uncertain Spatiotemporal, Contextual and Causal Inference*. Atlantis Thinking Machines. Atlantis Press, 2011. ISBN: 9789491216114. URL: https://books.google.com/books?id=g7UAIhnmJpsC.

[169] Ben Goertzel, Cassio Pennachin, and Nil Geisweiller. *Engineering General Intelligence, Part 1: A Path to Advanced AGI via Embodied Learning and Cognitive Synergy*. Atlantis Publishing Corporation, 2014. ISBN: 9789462390263.

[170] Irving John Good. "Speculations concerning the first ultraintelligent machine". In: *Advances in computers* 6.31 (1965), p. 88.

[171] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedyu. "Explaining and harnessing adversarial examples". In: Published as a conference paper at ICLR 2015. 2015. URL: https://pdfs.semanticscholar.org/bee0/44c8e8903fb67523c1f8c105ab4718600cdb.pdf.

[172] Ian J. Goodfellow et al. "Generative Adversarial Nets". In: *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*. 2014, pp. 2672–2680. URL: http://papers.nips.cc/paper/5423-generative-adversarial-nets.

[173] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. "Deep Learning, Chapters 6-12". Book in preparation for MIT Press. 2016. URL: http://www.deeplearningbook.org.

[174] Noah D Goodman and Joshua B. Tenenbaum. *Probabilistic Models of Cognition*. Accessed: 2016-11-15. 2016. URL: http://probmods.org/v2.

[175] Guido Governatori and Antonino Rotolo. "A defeasible logic of institutional agency". In: *IJCAI-03 Workshop on Nonmonotonic Reasoning, Action and Change*. IJCAI. 2003, pp. 97–104. URL: http://espace.library.uq.edu.au/view/UQ:9873/nrac03.pdf.

[176] Katja Grace. *Predictions of Human-Level AI Timelines*. Blog Post. 2015. URL: http://aiimpacts.org/predictions-of-human-level-ai-timelines/.

[177] Alex Graves, Greg Wayne, and Ivo Danihelka. "Neural turing machines". In: *arXiv preprint arXiv:1410.5401* (2014). URL: https://pdfs.semanticscholar.org/6eed/f0a4fe861335f7f7664c14de7f71c00b7932.pdf.

[178] Joshua D. Greene. "Beyond Point-and-Shoot Morality: Why Cognitive (Neuro)Science Matters for Ethics". In: *Ethics* 124.4 (2014), pp. 695–726. DOI: 10.1086/675875. eprint: http://dx.doi.org/10.1086/675875. URL: https://joshgreene.squarespace.com/s/beyond-point-and-shoot-morality-a4h2.pdf.

[179] Joshua D Greene. "Our driverless dilemma". In: *Science* 352.6293 (2016), pp. 1514–1515. URL: https://projects.iq.harvard.edu/files/mcl/files/greene-driverless-dilemma-sci16.pdf.

[180] Joshua Greene et al. "Embedding Ethical Principles in Collective Decision Support Systems". In: *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. AAAI'16. Phoenix, Arizona: AAAI Press, 2016, pp. 4147–4151. URL: http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/download/12457/12204.

[181] Karol Gregor et al. "DRAW: A Recurrent Neural Network For Image Generation". In: *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*. 2015, pp. 1462–1471. URL: http://jmlr.org/proceedings/papers/v37/gregor15.html.

[182] Yuhong Guo and Dale Schuurmans. "Convex Structure Learning for Bayesian Networks: Polynomial Feature Selection and Approximate Ordering". In: *Proceedings of the Twenty-Second Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*. AUAI Press, 2006, pp. 208–216. URL: https://webdocs.cs.ualberta.ca/~dale/papers/uai06.pdf.

[183] Sonal Gupta. "Distantly Supervised Information Extraction using Bootstrapped Patterns". PhD thesis. Stanford University, 2015. URL: http://nlp.stanford.edu/~manning/dissertations/Gupta-Sonal-thesis-augmented.pdf.

[184] Mordechai Guri et al. "Fansmitter: Acoustic Data Exfiltration from (Speakerless) Air-Gapped Computers". In: *CoRR* abs/1606.05915 (2016). URL: http://arxiv.org/abs/1606.05915.

[185] Dylan Hadfield-Menell et al. "Cooperative Inverse Reinforcement Learning". In: *CoRR* abs/1606.03137 (2016). URL: https://people.eecs.berkeley.edu/~dhm/papers/CIRL_NIPS_16.pdf.

[186] Dylan Hadfield-Menell et al. “The off-switch game”. In: abs/1611.08219 (2016). URL: https://arxiv.org/pdf/1611.08219v1.

[187] Jonathan Haidt. *The Righteous Mind: Why Good People Are Divided by Politics and Religion.* Random House, 2012.

[188] J. Storrs Hall. “Ethics for Self-Improving Machines”. In: *Machine Ethics.* Ed. by Michael Anderson and Susan L. Anderson. Cambridge University Press, 2011.

[189] Joseph Y. Halpern. *Reasoning About Uncertainty.* MIT Press, 2003. ISBN: 0262083205.

[190] Joseph Y Halpern and Rafael Pass. “Game theory with translucent players”. In: *arXiv preprint arXiv:1308.3778* (2013). URL: http://www.tark.org/proceedings/tark_jan7_13/p216-halpern.pdf.

[191] Joseph Y Halpern and Rafael Pass. “I don’t want to think about it now: Decision theory with costly computation”. In: *arXiv preprint arXiv:1106.2657* (2011). URL: http://www.cs.cornell.edu/~rafael/papers/compdec.pdf.

[192] Joseph Y Halpern, Rafael Pass, and Lior Seeman. “Decision Theory with Resource-Bounded Agents”. In: *Topics in cognitive science* 6.2 (2014), pp. 245–257. URL: http://www.cs.cornell.edu/home/halpern/papers/rbdec.pdf.

[193] Kristian J Hammond, Timothy M Converse, and Joshua W Grass. “The stabilization of environments”. In: *Artificial Intelligence* 72.1 (1995), pp. 305–327. URL: http://ac.els-cdn.com/000437029400006M/1-s2.0-000437029400006M-main.pdf?_tid=9424cbbc-c4e4-11e6-b2ae-00000aab0f6c&acdnat=1482039849_945550ec80cbdf0e7fd039ef161a2f13.

[194] Steve Hanneke. “A Bound on the Label Complexity of Agnostic Active Learning”. In: *Proceedings of the 24th International Conference on Machine Learning.* ICML ’07. ACM, 2007, pp. 353–360. ISBN: 978-1-59593-793-3. DOI: 10.1145/1273496.1273541. URL: http://doi.acm.org/10.1145/1273496.1273541.

[195] Steve Hanneke. “Theory of Disagreement-Based Active Learning”. In: *Foundations and Trends® in Machine Learning* 7.2-3 (2014), pp. 131–309. ISSN: 1935-8237. DOI: 10.1561/2200000037. URL: http://dx.doi.org/10.1561/2200000037.

[196] Lars Hansen. “Large Sample Properties of Generalized Method of Moments Estimators”. In: *Econometrica* 50.4 (1982), pp. 1029–54. URL: http://larspeterhansen.org/wp-content/uploads/2016/11/Hansen-econometrica-GMM.pdf.

[197] Lars Hansen. “Nobel Lecture: Uncertainty Outside and Inside Economic Models”. In: *Journal of Political Economy* 122.5 (2014), pp. 945 –987. URL: http://larspeterhansen.org/wp-content/uploads/2016/10/Uncertainty-Outside-and-Inside-Economic-Models.pdf.

[198] Ioannis Hatzilygeroudis and Jim Prentzas. “Neuro-Symbolic Approaches for Knowledge Representation in Expert Systems”. In: *Int. J. Hybrid Intell. Syst.* 1.3,4 (Dec. 2004), pp. 111–126. ISSN: 1448-5869. URL: https://ai2-s2-pdfs.s3.amazonaws.com/a2b0/1b08e1ebbc2d42399e58282d615be9bb97ee.pdf.

[199] Mark Herbster and Manfred K. Warmuth. “Tracking the Best Linear Predictor”. In: *The Journal of Machine Learning Research* 1 (Sept. 2001), pp. 281–309. ISSN: 1532-4435. DOI: 10.1162/153244301753683726. URL: http://dx.doi.org/10.1162/153244301753683726.

[200] Henry Hexmoor, Brian McLaughlan, and Gaurav Tuli. “Natural human role in supervising complex control systems”. In: *Journal of Experimental & Theoretical Artificial Intelligence* 21.1 (2009), pp. 59–77. URL: http://www2.cs.siu.edu/~hexmoor/CV/PUBLICATIONS/JOURNALS/JETAI-08/final.pdf.

[201] Bill Hibbard. *Model-based utility functions.* Vol. 3. 1. 2012, pp. 1–24. ISBN: 978-3-642-22887-2. DOI: 10.2478/v10229-011-0013-5. URL: https://pdfs.semanticscholar.org/6b2b/f1efaa66c77677070f1c52701f0f7f2a3e15.pdf.

[202] G. Hinton, O. Vinyals, and J. Dean. “Distilling the Knowledge in a Neural Network”. In: *ArXiv e-prints* (Mar. 2015). arXiv: 1503.02531 [stat.ML]. URL: https://arxiv.org/pdf/1503.02531v1.

[203] Geoffrey Hinton and Ruslan Salakhutdinov. “Reducing the Dimensionality of Data with Neural Networks”. In: *Science* 313.5786 (2006), pp. 504 –507. URL: https://www.cs.toronto.edu/~hinton/science.pdf.

[204] Daniel Hintze. “Problem Class Dominance in Predictive Dilemmas”. Honors Thesis. Arizona State University, 2014. URL: https://intelligence.org/wp-content/uploads/2014/10/Hintze-Problem-Class-Dominance-In-Predictive-Dilemmas.pdf.

[205] Jonathan Ho and Stefano Ermon. “Generative Adversarial Imitation Learning”. In: *CoRR* abs/1606.03476 (2016). URL: http://papers.nips.cc/paper/6391-generative-adversarial-imitation-learning.pdf.

[206] Thomas Hofmann, Bernhard Schölkopf, and Alexander J. Smola. “Kernel methods in machine learning”. In: *Annals of Statistics* 36.3 (2008), pp. 1171–1220. URL: http://www.kernel-machines.org/publications/pdfs/0701907.pdf.

[207] Eric J Horvitz. “Reasoning about beliefs and actions under computational resource constraints”. In: *Third AAAI Workshop on Uncertainty in Artificial Intelligence.* 1987, pp. 429–444. URL: ftp://131.107.65.22/pub/ejh/u87.pdf.

[208] Eric Horvitz. *One-Hundred Year Study of Artificial Intelligence: Reflections and Framing.* White paper. Stanford University, 2014. URL: https://ai100.stanford.edu/sites/default/files/ai100_framing_memo_0.pdf.

[209] Eric Horvitz and Bart Selman. *Interim Report from the Panel Chairs.* AAAI Presidential Panel on Long Term AI Futures. 2009. URL: https://www.aaai.org/Organization/Panel/panel-note.pdf.

[210] Lun-Kai Hsu, Tudor Achim, and Stefano Ermon. “Tight Variational Bounds via Random Projections and I-Projections”. In:

[211] Xueheng Hu. “Semantic Similarity in the Evaluation of Ontology Alignment”. PhD thesis. Miami University, 2011. URL: https://etd.ohiolink.edu/!etd.send_file?accession=miami1323323230&disposition=inline.

[212] Ling Huang et al. “Adversarial Machine Learning”. In: *Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence.* AISec ’11. ACM, 2011, pp. 43–58. ISBN: 978-1-4503-1003-1. DOI: 10.1145/2046684.2046692. URL: http://doi.acm.org/10.1145/2046684.2046692.

[213] Marcus Hutter. “One Decade of Universal Artificial Intelligence”. In: *CoRR* abs/1202.6153 (2012). URL: https://arxiv.org/pdf/1202.6153v1.

[214] Marcus Hutter. “Universal Algorithmic Intelligence: A mathematical top->down approach”. In: *CoRR* abs/cs/0701125 (2007). URL: https://arxiv.org/pdf/cs/0701125v1.

[215] Marcus Hutter et al. “Probabilities on Sentences in an Expressive Logic”. In: *Journal of Applied Logic* 11.4 (2013). Combining Probability and Logic: Papers from Progic 2011, pp. 386–420. ISSN: 1570-8683. DOI: 10.1016/j.jal.2013.03.003. URL: http://www.hutter1.net/publ/sproblogic.pdf.

[216] Tatyana Ivanova. “Ontology Alignment: State of the Art, Main Trends”. In: *Systems Approaches to Knowledge Management, Transfer, and Resource Development* (2012), p. 147.

[217] Garud N. Iyengar. “Robust Dynamic Programming”. In: *Mathematics of Operations Research* 30.2 (2005), pp. 257–280. DOI: 10.1287/moor.1040.0129. URL: http://dx.doi.org/10.1287/moor.1040.0129.

[218] Ariel Jaffe, Boaz Nadler, and Yuval Kluger. “Estimating the accuracies of multiple classifiers without labeled data”. In: *AISTATS.* Vol. 2. 2015, p. 4. URL: http://www.jmlr.org/proceedings/papers/v38/jaffe15.pdf.

[219] Dominik Janzing et al. “Quantifying causal influences”. In: *The Annals of Statistics* 41.5 (Oct. 2013), pp. 2324–2358. DOI: 10.1214/13-AOS1145. URL: http://dx.doi.org/10.1214/13-AOS1145.

[220] Julian Jara-Ettinger, Joshua B. Tenenbaum, and Laura Schulz. “Not so innocent: Reasoning about costs, competence, and culpability in very early childhood”. In: *Proceedings of the 35th Annual Meeting of the Cognitive Science Society, CogSci 2013, Berlin, Germany, July 31 - August 3, 2013.* 2013. URL: https://mindmodeling.org/cogsci2013/papers/0141/index.html.

[221] Julian Jara-Ettinger et al. “The Naive Utility Calculus: Computational Principles Underlying Commonsense Psychology”. In: *Trends in Cognitive Sciences* 20.8 (2016), pp. 589–604. URL: http://www.cell.com/trends/cognitive-sciences/pdf/S1364-6613(16)30124-3.pdf.

[222] Jean-Baptiste Jeannin et al. “A Formally Verified Hybrid System for the Next-Generation Airborne Collision Avoidance System”. In: *Tools and Algorithms for the Construction and Analysis of Systems: 21st International Conference, TACAS 2015, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2015, London, UK, April 11-18, 2015, Proceedings.* Ed. by Christel Baier and Cesare Tinelli. Springer Berlin Heidelberg, 2015, pp. 21–36. ISBN: 978-3-662-46681-0. DOI: 10.1007/978-3-662-46681-0_2. URL: http://www.cs.cmu.edu/~jeannin/papers/acasx.pdf.

[223] Taylor Jessica. *Informed oversight through an entropy-maximization objective.* Intelligent Agent Foundations Forum blog post. 2016. URL: https://agentfoundations.org/item?id=700.

[224] Taylor Jessica and Chris Olah. *Maximizing a quantity while ignoring effect through some channel.* Intelligent Agent Foundations Forum blog post. 2016. URL: https://agentfoundations.org/item?id=735.

[225] Fredrik D Johansson, Uri Shalit, and David Sontag. “Learning Representations for Counterfactual Inference”. In: *arXiv preprint arXiv:1605.03661* (2016). URL: https://arxiv.org/pdf/1605.03661.pdf.

[226] Kshitij Judah et al. “Active Imitation Learning: Formal and Practical Reductions to I.I.D. Learning”. In: *Journal of Machine Learning Research* 15 (2014), pp. 4105–4143. URL: http://jmlr.org/papers/v15/judah14a.html.

[227] Lukasz Kaiser and Ilya Sutskever. “Neural GPUs Learn Algorithms”. In: *CoRR* abs/1511.08228 (2015). URL: http://arxiv.org/abs/1511.08228.

[228] Lukasz Kaiser and Ilya Sutskever. “Neural GPUs Learn Algorithms”. In: *CoRR* abs/1511.08228 (2015). URL: https://arxiv.org/pdf/1511.08228v3.

[229] Holden Karnofsky. *Potential Risks from Advanced Artificial Intelligence: The Philanthropic Opportunity.* Blog Post. 2016. URL: http://www.openphilanthropy.org/blog/potential-risks-advanced-artificial-intelligence-philanthropic-opportunity.

[230] Andrej Karpathy. *Visualizing what ConvNets learn.* Blog Posts. 2015. URL: http://cs231n.github.io/understanding-cnn/.

[231] Fereshte Khani, Martin C. Rinard, and Percy Liang. “Unanimous Prediction for 100% Precision with Application to Learning Semantic Mappings”. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers.* 2016. URL: http://aclweb.org/anthology/P/P16/P16-1090.pdf.

[232] Carolyn Kim, Ashish Sabharwal, and Stefano Ermon. “Exact Sampling with Integer Linear Programs and Random Perturbations”. In: *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence.* AAAI’16. AAAI Press, 2016, pp. 3248–3254. URL: https://cs.stanford.edu/~ermon/papers/kim-sabharwal-ermon.pdf.

[233] Diederik P. Kingma and Max Welling. “Auto-Encoding Variational Bayes”. In: *ArXiv e-prints* (Dec. 2013). arXiv: 1312.6114 [stat.ML].

[234] J. Kirkpatrick et al. “Overcoming catastrophic forgetting in neural networks”. In: *ArXiv e-prints* (Dec. 2016). arXiv: 1612.00796 [cs.LG]. URL: https://arxiv.org/pdf/1612.00796v1.

[235] Gerwin Klein et al. “seL4: Formal verification of an OS kernel”. In: *Proceedings of the ACM SIGOPS 22nd symposium on Operating systems principles.* ACM. 2009, pp. 207–220. URL: http://web1.cs.columbia.edu/~junfeng/09fa-e6998/papers/sel4.pdf.

[236] Daniel Polani Klyubin Alexander S and Chrystopher L Nehaniv. “Empowerment: a universal agent-centric measure of control”. In: *Procs of the 2005 IEEE Congress on Evolutionary Computation 1 pp.128* - 135 (2005), pp. 128–135. URL: http://homepages.herts.ac.uk/~comqdp1/publications/files/cec2005_klyubin_polani_nehaniv.pdf.

[237] W. Bradley Knox and Peter Stone. “Interactively Shaping Agents via Human Reinforcement: The TAMER Framework”. In: *Proceedings of the Fifth International Conference on Knowledge Capture.* K-CAP ’09. Redondo Beach, California, USA: ACM, 2009, pp. 9–16. ISBN: 978-1-60558-658-8. DOI: 10.1145/1597735.1597738. URL: http://www.cs.utexas.edu/~pstone/Papers/bib2html-links/KCAP09-knox.pdf.

[238] Andrew J. Ko and Brad A. Myers. "Designing the Whyline: A Debugging Interface for Asking Questions About Program Behavior". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '04. Vienna, Austria: ACM, 2004, pp. 151–158. ISBN: 1-58113-702-8. DOI: 10.1145/985692.985712. URL: http://www.cs.cmu.edu/~ajko/papers/Ko2004Whyline.pdf.

[239] Andrew J. Ko and Brad A. Myers. "Finding Causes of Program Output with the Java Whyline". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '09. ACM, 2009, pp. 1569–1578. ISBN: 978-1-60558-246-7. DOI: 10.1145/1518701.1518942. URL: http://repository.cmu.edu/cgi/viewcontent.cgi?article=1163&context=hcii.

[240] Vadim Kosoy. "Optimal Polynomial-Time Estimators: A Bayesian Notion of Approximation Algorithm". In: *CoRR* abs/1608.04112 (2016). URL: https://arxiv.org/pdf/1608.04112v4.

[241] V. Krakovna and F. Doshi-Velez. "Increasing the Interpretability of Recurrent Neural Networks Using Hidden Markov Models". In: *ArXiv e-prints* (June 2016). arXiv: 1606.05320 [stat.ML]. URL: https://pdfs.semanticscholar.org/d512/e36d361d313cae20c9766fedd6f84c71f09c.pdf.

[242] Volodymyr Kuleshov and Percy S Liang. "Calibrated Structured Prediction". In: *Advances in Neural Information Processing Systems 28*. Ed. by C. Cortes et al. Curran Associates, Inc., 2015, pp. 3456–3464. URL: http://papers.nips.cc/paper/5658-calibrated-structured-prediction.pdf.

[243] Todd Kulesza et al. "Principles of Explanatory Debugging to Personalize Interactive Machine Learning". In: *Proceedings of the 20th International Conference on Intelligent User Interfaces*. IUI '15. Atlanta, Georgia, USA: ACM, 2015, pp. 126–137. ISBN: 978-1-4503-3306-1. DOI: 10.1145/2678025.2701399. URL: ftp://ftp.cs.orst.edu/pub/burnett/iui15-elucidebug.pdf.

[244] Tejas D. Kulkarni et al. "Hierarchical Deep Reinforcement Learning: Integrating Temporal Abstraction and Intrinsic Motivation". In: *CoRR* abs/1604.06057 (2016). URL: http://arxiv.org/abs/1604.06057.

[245] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. "Adversarial examples in the physical world". In: *CoRR* abs/1607.02533 (2016). URL: https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/45471.pdf.

[246] Patrick LaVictoire. *Proposal: Modeling goal stability in machine learning*. Blog post. 2015. URL: https://agentfoundations.org/item?id=130.

[247] Patrick LaVictoire et al. "Program Equilibrium in the Prisoner's Dilemma via Löb's Theorem". In: *AAAI Multiagent Interaction without Prior Coordination workshop*. 2014. URL: http://www.aaai.org/ocs/index.php/WS/AAAIW14/paper/viewFile/8833/8294.

[248] Brenden M. Lake, Ruslan Salakhutdinov, and Joshua B. Tenenbaum. "Human-level concept learning through probabilistic program induction". In: *Science* (2015). URL: http://web.mit.edu/cocosci/Papers/Science-2015-Lake-1332-8.pdf.

[249] Brenden M Lake et al. "Building Machines That Learn and Think like People Building Machines That Learn and Think like People". In: Open call for commentary proposals (until Nov. 22, 2016). 2016. URL: http://www.mit.edu/~tomeru/papers/machines_that_think.pdf.

[250] Axel van Lamsweerde. "Formal Specification: A Roadmap". In: *Proceedings of the Conference on The Future of Software Engineering*. ICSE '00. Limerick, Ireland: ACM, 2000, pp. 147–159. ISBN: 1-58113-253-0. DOI: 10.1145/336512.336546. URL: http://doi.acm.org/10.1145/336512.336546.

[251] H. Chad Lane et al. "Explainable Artificial Intelligence for Training and Tutoring". In: *Proceedings of the 12th International Conference on Artificial Intelligence in Education. Amersterdam, Holland: International Artificial Intelligence in Education Society*. July 2005. URL: http://ict.usc.edu/pubs/Explainable%20Artificial%20Intelligence%20for%20Training%20and%20Tutoring.pdf.

[252] Terran D Lane. "Machine learning techniques for the computer security domain of anomaly detection". PhD thesis. Purdue University, 2000. URL: https://www.cerias.purdue.edu/assets/pdf/bibtex_archive/98-11.pdf.

[253] M. Laskey et al. “Comparing Human-Centric and Robot-Centric Sampling for Robot Deep Learning from Demonstrations”. In: *ArXiv e-prints* (Oct. 2016). arXiv: 1610.00850 [cs.RO]. URL: https://arxiv.org/pdf/1610.00850.pdf.

[254] Alexander Lavin. “A Pareto Optimal D* Search Algorithm for Multiobjective Path Planning”. In: *CoRR* abs/1511.00787 (2015). URL: http://arxiv.org/abs/1511.00787.

[255] Neil Lawrence. *Discussion of ’Superintelligence: Paths, Dangers, Strategies’*. Discussion. 2016.

[256] Shane Legg and Marcus Hutter. “Universal Intelligence: A Definition of Machine Intelligence”. In: *Minds and Machines* 17.4 (2007), pp. 391–444. ISSN: 1572-8641. DOI: 10.1007/s11023-007-9079-x. URL: http://dx.doi.org/10.1007/s11023-007-9079-x.

[257] Jan Leike. *Nonparametric General Reinforcement Learning*. Tech. rep. 2016. URL: https://jan.leike.name/publications/Nonparametric%20General%20Reinforcement%20Learning%20-%20Leike%202016.pdf.

[258] Jan Leike, Jessica Taylor, and Benya Fallenstein. “A Formal Solution to the Grain of Truth Problem”. In: *CoRR* abs/1609.05058 (2016). URL: http://arxiv.org/abs/1609.05058.

[259] Michael van Lent, William Fisher, and Michael Mancuso. “An Explainable Artificial Intelligence System for Small-unit Tactical Behavior”. In: *Proceedings of the 16th Conference on Innovative Applications of Artifical Intelligence*. IAAI’04. AAAI Press, 2004, pp. 900–907. ISBN: 0-262-51183-5. URL: http://ict.usc.edu/pubs/An%20Explainable%20Artificial%20Intelligence%20System%20for%20Small-unit%20Tactical%20Behavior.pdf.

[260] Joshua Letchford, Vincent Conitzer, and Kamal Jain. “An ”Ethical” Game-Theoretic Solution Concept for Two-Player Perfect-Information Games”. In: *Internet and Network Economics, 4th International Workshop, WINE 2008, Shanghai, China, December 17-20, 2008. Proceedings*. 2008, pp. 696–707. DOI: 10.1007/978-3-540-92185-1_75. URL: http://dx.doi.org/10.1007/978-3-540-92185-1_75.

[261] Benjamin Letham et al. “Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model”. In: *The Annals of Applied Statistics* 9.3 (Sept. 2015), pp. 1350–1371. DOI: 10.1214/15-AOAS848. URL: http://dx.doi.org/10.1214/15-AOAS848.

[262] Guangliang Li et al. “Using Informative Behavior to Increase Engagement in the Tamer Framework”. In: *Proceedings of the 2013 International Conference on Autonomous Agents and Multi-agent Systems*. AAMAS ’13. International Foundation for Autonomous Agents and Multiagent Systems, 2013, pp. 909–916. ISBN: 978-1-4503-1993-5. URL: http://dl.acm.org/citation.cfm?id=2484920.2485064.

[263] Lihong Li, Michael L. Littman, and Thomas Walsh. “Knows what it knows: a framework for self-aware learning”. In: *Maching Learning* (2008). URL: http://icml2008.cs.helsinki.fi/papers/627.pdf.

[264] Longmei Li et al. “An Ontology of Preference-Based Multiobjective Evolutionary Algorithms”. In: *CoRR* abs/1609.08082 (2016). URL: http://arxiv.org/abs/1609.08082.

[265] Y. F. Li and Z. H. Zhou. “Towards Making Unlabeled Data Never Hurt”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.1 (2015), pp. 175–188. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2014.2299812.

[266] Yaliang Li et al. “A Survey on Truth Discovery”. In: *CoRR* abs/1505.02463 (2015). URL: https://arxiv.org/pdf/1505.02463v2.

[267] Percy Liang. *On the Elusiveness of a Specification for AI*. Presentation in (NIPS 2015) Symposium: Algorithms Among Us. 2015. URL: https://www.microsoft.com/en-us/research/video/symposium-algorithms-among-us-percy-liang/.

[268] Percy Liang and Dan Klein. “Analyzing the Errors of Unsupervised Learning”. In: *Proceedings of ACL-08: HLT*. Association for Computational Linguistics, 2008, pp. 879–887. URL: http://www.aclweb.org/anthology/P08-1100.

[269] Falk Lieder et al. “Algorithm selection by rational metareasoning as a model of human strategy selection”. In: *Advances in Neural Information Processing Systems*. MIT Press, 2014, pp. 2870–2878. URL: http://papers.nips.cc/paper/5552-global-sensitivity-analysis-for-map-inference-in-graphical-models.pdf.

[270] H. W. Lin and M. Tegmark. "Why does deep and cheap learning work so well?" In: *ArXiv e-prints* (Aug. 2016). arXiv: 1608.08225 [cond-mat.dis-nn]. URL: https://arxiv.org/pdf/1608.08225v1.pdf.

[271] Xiao Lin and Devi Parikh. "Don't Just Listen, Use Your Imagination: Leveraging Visual Common Sense for Non-Visual Tasks". In: *CoRR* abs/1502.06108 (2015). URL: https://arxiv.org/pdf/1502.06108v3.

[272] Zachary C. Lipton et al. "Combating Reinforcement Learning's Sisyphean Curse with Intrinsic Fear". In: *CoRR* abs/1611.01211 (2016). URL: https://arxiv.org/pdf/1611.01211v3.

[273] Zachary Chase Lipton. "The Mythos of Model Interpretability". In: *CoRR* abs/1606.03490 (2016). URL: http://zacklipton.com/media/papers/mythos-model-interpretability_16.pdf.

[274] C. Liu, X. Xu, and D. Hu. "Multiobjective Reinforcement Learning: A Comprehensive Overview". In: *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 45.3 (2015), pp. 385–398. ISSN: 2168-2216. DOI: 10.1109/TSMC.2014.2358639. URL: http://ieeexplore.ieee.org/document/6918520/.

[275] Chang Liu et al. "Goal Inference Improves Objective and Perceived Performance in Human-Robot Collaboration". In: *Proceedings of the 2016 International Conference on Autonomous Agents &#38; Multiagent Systems*. AAMAS '16. International Foundation for Autonomous Agents and Multiagent Systems, 2016, pp. 940–948. ISBN: 978-1-4503-4239-1. URL: http://www.jesshamrick.com/publications/pdf/Liu2016-Goal_Inference_Improves_Objective.pdf.

[276] H. Liu and H. Motoda. *Computational Methods of Feature Selection*. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. CRC Press, 2007. ISBN: 9781584888796. URL: https://books.google.co.in/books?id=N1ViHNWZeQ0C.

[277] Sarah M. Loos, David Renshaw, and Andre Platzer. "Formal Verification of Distributed Aircraft Controllers". In: *Proceedings of the 16th International Conference on Hybrid Systems: Computation and Control*. HSCC '13. ACM, 2013, pp. 125–130. ISBN: 978-1-4503-1567-8. DOI: 10.1145/2461328.2461350. URL: http://symbolaris.com/pub/discworld.pdf.

[278] Georgios Loukas. "Defence against denial of service in self-aware networks". PhD thesis. Imperial College, 2006. URL: https://pdfs.semanticscholar.org/c328/80da0cf2e729e125f9e3034664f83e3b686c.pdf.

[279] John Lygeros, Claire Tomlin, and Shankar Sastry. "Controllers for Reachability Specifications for Hybrid Systems". In: *Automatica* 35.3 (Mar. 1999), pp. 349–370. ISSN: 0005-1098. URL: http://dx.doi.org/10.1016/S0005-1098(98)00193-9.

[280] Laurens Van Der Maaten and Hinton Geoffrey E. "Visualizing High-Dimensional Data Using t-SNE". In: *Journal of Machine Learning Research* 9 (2008), pp. 2579–2605.

[281] William MacAskill. "Normative Uncertainty". PhD thesis. University of Oxford, 2014. URL: http://commonsenseatheism.com/wp-content/uploads/2014/03/MacAskill-Normative-Uncertainty.pdf.

[282] Assia Mahboubi and Enrico Tassi. "The Mathematical Components Library: Principles and Design Choices". presentation. 2013. URL: http://ssr.msr-inria.inria.fr/doc/tutorial-itp13/slides.pdf.

[283] Aravindh Mahendran and Andrea Vedaldi. "Understanding Deep Image Representations by Inverting Them". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015. URL: https://www.robots.ox.ac.uk/~vedaldi/assets/pubs/mahendran15understanding.pdf.

[284] Richard Mallah. *Approaching Value Alignment*. Forthcoming. 2017.

[285] Richard Mallah. *Meaning, Leverage, and the Future*. Presentation. 2014. URL: https://people.eecs.berkeley.edu/~russell/research/future/mallah-aamas14-future.pdf.

[286] Richard Mallah. *Ontology Alignment in Business and In Life*. Presented at Smart Data Conference, San Jose, California, August 19. 2015. URL: http://smartdata2015.dataversity.net/sessionPop.cfm?confid=91&proposalid=7754.

[287] Richard Mallah. "The Top A.I. Breakthroughs of 2015". In: (2015). KurzweilAINetwork Blog post. URL: http://www.kurzweilai.net/the-top-ai-breakthroughs-of-2015.

[288] David Manheim. *Goodhart's Law and Why Measurement is Hard*. Blog Posts. 2016. URL: http://www.ribbonfarm.com/2016/06/09/goodharts-law-and-why-measurement-is-hard/.

[289] Daniel Mankowitz, Aviv Tamar, and Shie Mannor. *Situational Awareness by Risk-Conscious Skills*. Tech. rep. 2016. URL: https://4caf2f9f-a-62cb3a1a-s-sites.googlegroups.com/site/wildml2016/mankowitz16saricos.pdf.

[290] Gideon S. Mann and Andrew McCallum. "Generalized Expectation Criteria for Semi-Supervised Learning with Weakly Labeled Data". In: *The Journal of Machine Learning Research* 11 (Mar. 2010), pp. 955–984. ISSN: 1532-4435. URL: http://www.jmlr.org/papers/volume11/mann10a/mann10a.pdf.

[291] David Martens and Foster Provost. "Explaining Data-driven Document Classifications". In: *MIS Q.* 38.1 (Mar. 2014), pp. 73–100. ISSN: 0276-7783. URL: pages.stern.nyu.edu/%7Efprovost/Papers/MartensProvost_Explaining.pdf.

[292] Bernard Merialdo. "Tagging English Text with a Probabilistic Model". In: *Computational Linguistics* 20.2 (June 1994), pp. 155–171. ISSN: 0891-2017. URL: http://aclweb.org/anthology/J94-2001.pdf.

[293] Tomas Mikolov, Armand Joulin, and Marco Baroni. "A Roadmap towards Machine Intelligence". In: *CoRR* abs/1511.08130 (2015). URL: https://pdfs.semanticscholar.org/8097/95994cfaf0ce464848c99816b16
pdf.

[294] Mike Mintz et al. "Distant Supervision for Relation Extraction Without Labeled Data". In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*. ACL '09. Association for Computational Linguistics, 2009, pp. 1003–1011. ISBN: 978-1-932432-46-6. URL: https://web.stanford.edu/~jurafsky/mintz.pdf.

[295] Ian M. Mitchell, Alexandre M. Bayen, and Claire J Tomlin. "A time-dependent Hamilton-Jacobi formulation of reachable sets for continuous dynamic games". In: *IEEE Transactions on Automatic Control* 50.7 (2005), pp. 947–957. ISSN: 0018-9286. DOI: 10.1109/TAC.2005.851439. URL: https://www.cs.ubc.ca/~mitchell/Papers/publishedIEEEtac05.pdf.

[296] Stefan Mitsch, Sarah M. Loos, and Andre Platzer. "Towards Formal Verification of Freeway Traffic Control". In: *Proceedings of the 2012 IEEE/ACM Third International Conference on Cyber-Physical Systems*. ICCPS '12. IEEE Computer Society, 2012, pp. 171–180. ISBN: 978-0-7695-4695-7. URL: http://dx.doi.org/10.1109/ICCPS.2012.25.

[297] Stefan Mitsch et al. "Formal Verification of Obstacle Avoidance and Navigation of Ground Robots". In: *CoRR* abs/1605.00604 (2016). URL: http://arxiv.org/abs/1605.00604.

[298] Shakir Mohamed and Danilo J. Rezende. "Variational Information Maximisation for Intrinsically Motivated Reinforcement Learning". In: *Proceedings of the 28th International Conference on Neural Information Processing Systems*. NIPS'15. Montreal, Canada: MIT Press, 2015, pp. 2125–2133. URL: https://papers.nips.cc/paper/5668-variational-information-maximisation-for-intrinsically-motivated-reinforcement-learning.pdf.

[299] Teodor M. Moldovan and Pieter Abbeel. "Safe Exploration in Markov Decision Processes". In: *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*. Ed. by John Langford and Joelle Pineau. New York, NY, USA: ACM, 2012, pp. 1711–1718. URL: http://icml.cc/2012/papers/838.pdf.

[300] Seyed-Mohsen Moosavi-Dezfooli et al. "Universal adversarial perturbations". In: *CoRR* abs/1610.08401 (2016). URL: http://arxiv.org/abs/1610.08401.

[301] Alexander Mordvintsev, Christopher Olah, and Mike Tyka. *Inceptionism: Going deeper into neural networks*. Google Research Blog post. 2015. URL: https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html.

[302] Luke Muehlhauser. *Transparency in Safety-Critical Systems*. Machine Intelligence Research Institute, 2013. URL: https://intelligence.org/2013/08/25/transparency-in-safety-critical-systems/.

[303] Luke Muehlhauser and Bill Hibbard. "Exploratory Engineering in Artificial Intelligence". In: *Commun. ACM* 57.9 (Sept. 2014), pp. 32–34. ISSN: 0001-0782. DOI: 10.1145/2644257. URL: http://intelligence.org/files/ExploratoryEngineeringAI.pdf.

[304] Luke Muehlhauser and Anna Salamon. “Intelligence Explosion: Evidence and Import”. In: *Singularity Hypotheses: A Scientific and Philosophical Assessment.* Ed. by Amnon H. Eden et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 15–42. ISBN: 978-3-642-32560-1. DOI: 10.1007/978-3-642-32560-1_2. URL: https://intelligence.org/files/IE-EI.pdf.

[305] Luke Muehlhauser and Chris Williamson. *Ideal Advisor Theories and Personal CEV.* Machine Intelligence Research Institute, 2013. URL: https://intelligence.org/files/IdealAdvisorTheories.pdf.

[306] Jeanne-Marie Musca. *The Cyber Grand Challenge Autonomous detection and patching of online vulnerabilities.* 2016. URL: http://www.cs.tufts.edu/comp/116/archive/fall2016/jmusca.pdf.

[307] Gina Neff and Peter Nagy. “Automation, Algorithms, and Politics— Talking to Bots: Symbiotic Agency and the Case of Tay”. In: *International Journal of Communication* 10 (2016), p. 17. URL: http://ijoc.org/index.php/ijoc/article/viewFile/6277/1804.

[308] Jersey Neyman. *Sur les applications de la theorie des probabilities aux experiences agricoles: Essai des principes.* 1923.

[309] Andrew Y. Ng, Daishi Harada, and Stuart J. Russell. “Policy Invariance Under Reward Transformations: Theory and Application to Reward Shaping”. In: *Proceedings of the Sixteenth International Conference on Machine Learning.* ICML ’99. Morgan Kaufmann Publishers Inc., 1999, pp. 278–287. ISBN: 1-55860-612-2. URL: http://www.robotics.stanford.edu/~ang/papers/shaping-icml99.pdf.

[310] Andrew Y Ng and Stuart Russell. “Algorithms for Inverse Reinforcement Learning”. In: *in Proc. 17th International Conf. on Machine Learning.* Citeseer. 2000. URL: https://people.eecs.berkeley.edu/~russell/papers/ml00-irl.pdf.

[311] Anh Nguyen, Jason Yosinski, and Jeff Clune. “Deep neural networks are easily fooled: High confidence predictions for unrecognizable images”. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 2015, pp. 427–436. DOI: 10.1109/CVPR.2015.7298640. URL: http://yosinski.com/media/papers/Nguyen__2015__CVPR__Deep_Neural_Networks_Are_Easily_Fooled.pdf.

[312] Anh Nguyen et al. “Synthesizing the preferred inputs for neurons in neural networks via deep generator networks”. In: *Advances in Neural Information Processing Systems 29.* 2016. URL: https://papers.nips.cc/paper/6519-synthesizing-the-preferred-inputs-for-neurons-in-neural-networks-via-deep-generator-networks.pdf.

[313] Kamal Nigam et al. “Learning to Classify Text from Labeled and Unlabeled Documents”. In: *Proceedings of the Fifteenth National/Tenth Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence.* AAAI ’98/IAAI ’98. American Association for Artificial Intelligence, 1998, pp. 792–799. ISBN: 0-262-51098-7. URL: http://www.kamalnigam.com/papers/emcat-aaai98.pdf.

[314] D. Nikolić. “Practopoiesis: Or how life fosters a mind”. In: *ArXiv e-prints* (Feb. 2014). arXiv: 1402.5332 [q-bio.NC]. URL: http://ac.els-cdn.com/S002251931500106X/1-s2.0-S002251931500106X-main.pdf?_tid=6f1c2bc8-c4e7-11e6-972c-00000aacb35e&acdnat=1482041076_291e3fa094036bcad0f5601cfc975933.

[315] Arnab Nilim and Laurent El Ghaoui. “Robust Control of Markov Decision Processes with Uncertain Transition Matrices”. In: *Oper. Res.* 53.5 (Sept. 2005), pp. 780–798. ISSN: 0030-364X. DOI: 10.1287/opre.1050.0216. URL: http://dx.doi.org/10.1287/opre.1050.0216.

[316] Hiroki Nishimura and Efe A. Ok. “Utility representation of an incomplete and nontransitive preference relation”. In: *Journal of Economic Theory* 166 (2016), pp. 164 –185. ISSN: 0022-0531. DOI: http://dx.doi.org/10.1016/j.jet.2016.07.002. URL: http://hirokinishimura.net/files/BinRelRep.pdf.

[317] E Nivel et al. “Bounded Recursive Self-Improvement”. In: *arXiv preprint arXiv:1312.6764* (2013). URL: http://people.idsia.ch/~steunebrink/Publications/TR13_bounded_recursive_self-improvement.pdf.

[318] Natalya F. Noy and Michel Klein. “Ontology Evolution: Not the Same as Schema Evolution”. In: *Knowledge and Information Systems* 6.4 (2004), pp. 428–440. ISSN: 0219-3116. DOI: 10.1007/s10115-003-0137-2. URL: https://pdfs.semanticscholar.org/e38f/9d0878d9b06713142331695efe9ce5e5e0e0.pdf.

[319] E. Ohn-Bar and M. M. Trivedi. “Looking at Humans in the Age of Self-Driving and Highly Automated Vehicles”. In: *IEEE Transactions on Intelligent Vehicles* 1.1 (2016), pp. 90–104. ISSN: 2379-8858. DOI: 10.1109/TIV.2016.2571067. URL: http://cvrr.ucsd.edu/eshed/papers/humansTIV.pdf.

[320] Chris Olah. *Visualizing Representations: Deep Learning and Human Beings.* blog post. 2015. URL: http://colah.github.io/posts/2015-01-Visualizing-Representations/.

[321] Stephen M. Omohundro. “The Basic AI Drives”. In: *Proceedings of the 2008 Conference on Artificial General Intelligence 2008: Proceedings of the First AGI Conference.* IOS Press, 2008, pp. 483–492. ISBN: 978-1-58603-833-5. URL: https://selfawaresystems.files.wordpress.com/2008/01/ai_drives_final.pdf.

[322] Stephen M Omohundro. *The nature of self-improving artificial intelligence.* Presented at Singularity Summit 2007. 2007. URL: http://selfawaresystems.files.wordpress.com/2008/01/nature_of_self_improving_ai.pdf.

[323] OpenAI. *OpenAI Universe.* 2016. URL: https://universe.openai.com/.

[324] Toby Ord. “Moral Trade”. In: *Ethics* 126.1 (2015), pp. 118–138. DOI: 10.1086/682187. eprint: http://dx.doi.org/10.1086/682187. URL: http://amirrorclear.net/files/moral-trade.pdf.

[325] Toby Ord, Rafaela Hillerbrand, and Anders Sandberg. “Probing the improbable: methodological challenges for risks with low probabilities and high stakes”. In: *Journal of Risk Research* 13.2 (2010), pp. 191–205. DOI: 10.1080/13669870903126267. eprint: http://dx.doi.org/10.1080/13669870903126267. URL: http://www.amirrorclear.net/files/probing-the-improbable.pdf.

[326] Laurent Orseau and Stuart Armstrong. “Safely Interruptible Agents”. In: *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence, UAI 2016, June 25-29, 2016, New York City, NY, USA.* AUAI Press Corvallis, Oregon. 2016, pp. 557–566. URL: http://auai.org/uai2016/proceedings/papers/68.pdf.

[327] Laurent Orseau and Mark Ring. “Self-Modification and Mortality in Artificial Agents”. In: *Artificial General Intelligence: 4th International Conference, AGI 2011, Mountain View, CA, USA, August 3-6, 2011. Proceedings.* Ed. by Jürgen Schmidhuber, Kristinn R. Thórisson, and Moshe Looks. Springer Berlin Heidelberg, 2011, pp. 1–10. ISBN: 978-3-642-22887-2. DOI: 10.1007/978-3-642-22887-2_1. URL: http://dx.doi.org/10.1007/978-3-642-22887-2_1.

[328] Laurent Orseau and Mark Ring. “Space-Time embedded intelligence”. In: *Artificial General Intelligence.* Springer, 2012, pp. 209–218. URL: http://agi-conference.org/2012/wp-content/uploads/2012/12/paper_76.pdf.

[329] Ian Osband et al. “Deep Exploration via Bootstrapped DQN”. In: *Advances In Neural Information Processing Systems 29.* Ed. by D. D. Lee et al. Curran Associates, Inc., 2016, pp. 4026–4034. URL: http://papers.nips.cc/paper/6500-deep-exploration-via-bootstrapped-dqn.pdf.

[330] Martin Ouimet and Kristina Lundqvist. *Formal Software Verification: Model Checking and Theorem Proving.* Tech. rep. MIT Embedded Systems Laboratory, 2007. URL: https://pdfs.semanticscholar.org/c593/e5fc056b519cd43b5fdf033eb3281cd74983.pdf.

[331] Sinno Jialin Pan and Qiang Yang. “A Survey on Transfer Learning”. In: *IEEE Transactions on Knowledge and Data Engineering* 22.10 (2010), pp. 1345–1359. ISSN: 1041-4347. DOI: 10.1109/TKDE.2009.191. URL: https://www.cse.ust.hk/~qyang/Docs/2009/tkde_transfer_learning.pdf.

[332] Nicolas Papernot et al. “Practical Black-Box Attacks against Deep Learning Systems using Adversarial Examples”. In: *CoRR* abs/1602.02697 (2016). URL: http://arxiv.org/abs/1602.02697.

[333] Raja Parasuraman, Thomas B Sheridan, and Christopher D Wickens. “A model for types and levels of human interaction with automation”. In: *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on* 30.3 (2000), pp. 286–297. URL: https://hci.cs.uwaterloo.ca/faculty/elaw/cs889/reading/automation/sheridan.pdf.

[334] Emilio Parisotto et al. “Neuro-Symbolic Program Synthesis”. In: *CoRR* abs/1611.01855 (2016). URL: http://arxiv.org/abs/1611.01855.

[335] Judea Pearl. “Causal inference in statistics: An overview”. In: *Statistics Surveys* 3 (2009), pp. 96–146. DOI: 10.1214/09-SS057. URL: http://dx.doi.org/10.1214/09-SS057.

[336] Judea Pearl. *Causality: models, reasoning, and inference*. 1st ed. Cambridge University Press, 2000.

[337] Judea Pearl. *Causality: models, reasoning, and inference*. 2nd ed. Cambridge University Press, 2009.

[338] Martin Pecka and Tomas Svoboda. “Safe Exploration Techniques for Reinforcement Learning–An Overview”. In: *Modelling and Simulation for Autonomous Systems: First International Workshop, MESAS 2014, Rome, Italy, May 5-6, 2014, Revised Selected Papers*. Ed. by Jan Hodicky. Springer International Publishing, 2014, pp. 357–375. ISBN: 978-3-319-13823-7. DOI: 10.1007/978-3-319-13823-7_31. URL: http://cmp.felk.cvut.cz/~peckama2/papers/safe_exploration_overview_lncs.pdf.

[339] Vittorio Perera et al. “Dynamic generation and refinement of robot verbalization”. In: *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. 2016, pp. 212–218. DOI: 10.1109/ROMAN.2016.7745133. URL: http://www.cs.cmu.edu/~mmv/papers/16roman-verbalization.pdf.

[340] Jonas Peters et al. “Causal Discovery with Continuous Additive Noise Models”. In: *The Journal of Machine Learning Research* 15.1 (Jan. 2014), pp. 2009–2053. ISSN: 1532-4435. URL: http://jmlr.org/papers/volume15/peters14a/peters14a.pdf.

[341] Federico Pistono and Roman V. Yampolskiy. “Unethical Research: How to Create a Malevolent Artificial Intelligence”. In: *CoRR* abs/1605.02817 (2016). URL: https://arxiv.org/ftp/arxiv/papers/1605/1605.02817.pdf.

[342] Emmanouil Antonios Platanios, Avrim Blum, and Tom M Mitchell. “Estimating Accuracy from Unlabeled Data”. In: *Conference on Uncertainty in Artificial Intelligence*. 2014, pp. 1–10. URL: http://auai.org/uai2014/proceedings/individuals/313.pdf.

[343] Andr Platzer. *Logical analysis of hybrid systems: proving theorems for complex dynamics*. Springer Publishing Company, Incorporated, 2010. URL: https://www.isr.umd.edu/sites/default/files/110919_Platzer.pdf.

[344] Alexey Potapov and Sergey Rodionov. “Universal Empathy and Ethical Bias for Artificial General Intelligence”. In: *CoRR* abs/1308.0702 (2013). URL: https://arxiv.org/pdf/1308.0702v1.

[345] Alexey Potapov et al. “Cognitive Bias for Universal Algorithmic Intelligence”. In: *CoRR* abs/1209.4290 (2012). URL: https://arxiv.org/pdf/1209.4290v1.

[346] Walter W Powell and Laurel Smith-Doerr. “Networks and economic life”. In: *The handbook of economic sociology* 368 (1994), p. 380. URL: http://woodypowell.com/wp-content/uploads/2012/03/4_NetworksandEconomicLife.pdf.

[347] Bart Presnell, Ryan Houlette, and Dan Fu. “Making behavior modeling accessible to non-programmers: challenges and solutions”. In: *Interservice/Industry Training, Simulation, and Education Conference, Orlando, FL*. 2007. URL: https://simbionic.com/papers/IITSEC-07-behavior-modeling.pdf.

[348] Matthew J Probst and Sneha Kumar Kasera. “Statistical trust establishment in wireless sensor networks”. In: *Parallel and Distributed Systems, 2007 International Conference on*. Vol. 2. IEEE. 2007, pp. 1–8. URL: http://www.cs.utah.edu/~kasera/myPapers/trust.pdf.

[349] Luca Pulina and Armando Tacchella. “An abstraction-refinement approach to verification of artificial neural networks”. In: *Computer Aided Verification*. Springer. 2010, pp. 243–257. URL: https://pdfs.semanticscholar.org/72e5/5b90b5b791646266b0da8f6528d99aa96be5.pdf.

[350] Steve Rabin. *Introduction to game development*. 2nd ed. Nelson Education, 2010.

[351] Nathan D. Ratliff, J. Andrew Bagnell, and Martin A. Zinkevich. “Maximum Margin Planning”. In: *Proceedings of the 23rd International Conference on Machine Learning*. ICML ’06. ACM, 2006, pp. 729–736. ISBN: 1-59593-383-2. DOI: 10.1145/1143844.1143936. URL: http://martin.zinkevich.org/publications/maximummarginplanning.pdf.

[352] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “”Why Should I Trust You?”: Explaining the Predictions of Any Classifier”. In: *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’16. San Francisco, California, USA: ACM, 2016, pp. 1135–1144. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939778. URL: http://www.kdd.org/kdd2016/papers/files/rfp0573-ribeiroA.pdf.

[353] Konrad Rieck et al. “Automatic analysis of malware behavior using machine learning”. In: *Journal of Computer Security* 19.4 (2011), pp. 639–668. URL: http://www.mlsec.org/malheur/docs/malheur-jcs.pdf.

[354] Mark Ring and Laurent Orseau. “Delusion, Survival, and Intelligent Agents”. In: *Artificial General Intelligence: 4th International Conference, AGI 2011, Mountain View, CA, USA, August 3-6, 2011. Proceedings*. Ed. by Jürgen Schmidhuber, Kristinn R. Thórisson, and Moshe Looks. Springer Berlin Heidelberg, 2011, pp. 11–20. ISBN: 978-3-642-22887-2. DOI: 10.1007/978-3-642-22887-2_2. URL: http://people.idsia.ch/~ring/AGI-2011/Paper-B.pdf.

[355] RobbBB. “Building Phenomenological Bridges”. In: (2013).

[356] Marko Robnik-Sikonja and Igor Kononenko. “Explaining classifications for individual instances”. In: *Knowledge and Data Engineering, IEEE Transactions on* 20.5 (2008), pp. 589–600.

[357] Marek Rosa et al. *GoodAI Agent Development Roadmap*. 2016. URL: http://media.wix.com/ugd/2f0a43_091d76d2b0354b0db4d88c3a57fdf76d.pdf.

[358] Stéphane Ross, Geoffrey J. Gordon, and Drew Bagnell. “A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning”. In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS-11)*. Ed. by Geoffrey J. Gordon and David B. Dunson. Vol. 15. Journal of Machine Learning Research - Workshop and Conference Proceedings, 2011, pp. 627–635. URL: http://www.jmlr.org/proceedings/papers/v15/ross11a/ross11a.pdf.

[359] Francesca Rossi. “Ethical Preference-Based Decision Support Systems”. In: *LIPIcs-Leibniz International Proceedings in Informatics*. Vol. 59. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik. 2016. URL: http://drops.dagstuhl.de/opus/volltexte/2016/6187/pdf/LIPIcs-CONCUR-2016-2.pdf.

[360] Francesca Rossi. *Moral Preferences*. Presented at the 2016 Colloquium Series on Robust and Beneficial AI (CSRBAI). 2016. URL: https://intelligence.org/files/csrbai/pref-eth1.pdf.

[361] Damien Rouhling et al. “Axiomatic constraint systems for proof search modulo theories”. In: *10th International Symposium on Frontiers of Combining Systems (FroCoS’15)*. Ed. by C. Lutz and S. Ranise. Vol. 9322. LNAI. Wroclaw, Poland: Springer, Sept. 2015. DOI: 10.1007/978-3-319-24246-0_14. URL: https://hal.inria.fr/hal-01107944.

[362] Donald B. Rubin. “Estimating causal effects of treatments in randomized and nonrandomized studies”. In: *Journal of Educational Psychology* 66.5 (Oct. 1974), pp. 688–701. ISSN: 0022-0663. URL: http://dx.doi.org/10.1037/h0037350.

[363] Stuart J Russell and Devika Subramanian. “Provably bounded-optimal agents”. In: *Journal of Artificial Intelligence Research* (1995), pp. 1–36. URL: https://www.jair.org/media/133/live-133-1446-jair.pdf.

[364] Stuart J. Russell et al. “The Physics of Text: Ontological Realism in Information Extraction”. In: *Proceedings of the 5th Workshop on Automated Knowledge Base Construction, AKBC@NAACL-HLT 2016, San Diego, CA, USA, June 17, 2016*. 2016, pp. 51–56. URL: http://aclweb.org/anthology/W/W16/W16-1310.pdf.

[365] Stuart Russell. “Learning agents for uncertain environments”. In: *Proceedings of the eleventh annual conference on Computational learning theory*. ACM. 1998, pp. 101–103. URL: https://people.eecs.berkeley.edu/~russell/papers/colt98-uncertainty.pdf.

[366] Stuart Russell. “Moral Philosophy Will Become Part of the Tech Industry”. In: *Time Magazine* (2016).

[367] Stuart Russell. *Of Myths and Moonshine*. Conversation on Edge.org. 2014. URL: https://www.edge.org/conversation/themyth-of-ai#26015.

[368] Stuart Russell. “Should we fear super smart robots?” In: *Scientific American* 314.6 (2016), pp. 58–59. URL: https://people.eecs.berkeley.edu/~russell/papers/sciam16-supersmart.pdf.

[369] Stuart Russell. “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*. Ed. by Anne Laurent et al. Springer International Publishing, 2014, pp. 10–14. ISBN: 978-3-319-08795-5. DOI: 10.1007/978-3-319-08795-5_2. URL: http://dx.doi.org/10.1007/978-3-319-08795-5_2.

[370] Stuart Russell. “Unifying Logic and Probability”. In: *Communications of the ACM* 58.7 (July 2015), pp. 88–97. DOI: 10.1145/2699411. URL: https://people.eecs.berkeley.edu/~russell/papers/cacm15-oupm.pdf.

[371] Stuart Russell, Daniel Dewey, and Max Tegmark. “Research Priorities for Robust and Beneficial Artificial Intelligence”. In: *AI Magazine* 36.4 (2015). URL: http://futureoflife.org/data/documents/research_priorities.pdf.

[372] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach, Chapters 1-17*. Pearson, 2009.

[373] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. 3rd. Pearson, 2010.

[374] Andrei A. Rusu et al. “Progressive Neural Networks”. In: *CoRR* abs/1606.04671 (2016). URL: https://arxiv.org/pdf/1606.04671v3.

[375] Jordi Sabater and Carles Sierra. “Review on computational trust and reputation models”. In: *Artificial intelligence review* 24.1 (2005), pp. 33–60. URL: http://www.iiia.csic.es/files/pdfs/1035.pdf.

[376] Christoph Salge, Cornelius Glackin, and Daniel Polani. “Empowerment–An Introduction”. In: *Guided Self-Organization: Inception*. Ed. by Mikhail Prokopenko. Springer Berlin Heidelberg, 2014, pp. 67–114. ISBN: 978-3-642-53734-9. DOI: 10.1007/978-3-642-53734-9_4. URL: https://pdfs.semanticscholar.org/d01e/3414ca706eda917576d947ece811b5cbcdde.pdf.

[377] Anders Sandberg and Nick Bostrom. “Whole brain emulation: A Roadmap”. In: (2008). URL: http://www.philosophy.ox.ac.uk/__data/assets/pdf_file/0019/3853/brain-emulation-roadmap-report.pdf.

[378] J Denis Sargan. “The Estimation of Economic Relationships using Instrumental Variables”. In: *Econometrica* 26.3 (1958), pp. 393–415. ISSN: 00129682, 14680262.

[379] J Denis Sargan. “The Estimation of Relationships with Autocorrelated Residuals by the Use of Instrumental Variables”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 21.1 (1959), pp. 91–105. ISSN: 00359246.

[380] Gopal Sarma and Nick J. Hay. “Mammalian Value Systems”. In: *CoRR* abs/1607.08289 (2016). URL: http://arxiv.org/abs/1607.08289.

[381] Jens Schreiter et al. “Safe Exploration for Active Learning with Gaussian Processes”. In: *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2015, Porto, Portugal, September 7-11, 2015, Proceedings, Part III*. Cham: Springer International Publishing, 2015, pp. 133–149. ISBN: 978-3-319-23461-8. DOI: 10.1007/978-3-319-23461-8_9. URL: http://www.jmlr.org/proceedings/papers/v37/sui15.pdf.

[382] John Schulman et al. “High-Dimensional Continuous Control Using Generalized Advantage Estimation”. In: *Proceedings of the International Conference on Learning Representations (ICLR)*. 2016. URL: https://arxiv.org/pdf/1506.02438.pdf.

[383] Johann M Schumann and Yan Liu. *Applications of neural networks in high assurance systems*. Springer, 2010.

[384] D. Sculley et al. “Machine Learning: The High Interest Credit Card of Technical Debt”. In: *SE4ML: Software Engineering for Machine Learning (NIPS 2014 Workshop)*. 2014. URL: http://research.google.com/pubs/archive/43146.pdf.

[385] Bart Selman. *Scaling-up AI Systems: Insights From Computational Complexity.* forthcoming. 2017.

[386] Sanjit A. Seshia, Dorsa Sadigh, and S. Shankar Sastry. *Towards Verified Artificial Intelligence.* Tech. rep. EECS Department, University of California, Berkeley, 2016. URL: `https://people.eecs.berkeley.edu/~dsadigh/Papers/seshia-verifiedAI-arxiv.pdf`.

[387] Burr Settles. *Active learning literature survey.* Tech. rep. 2010. URL: `http://burrsettles.com/pub/settles.activelearning.pdf`.

[388] H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. "Query by Committee". In: *Proceedings of the Fifth Annual Workshop on Computational Learning Theory.* COLT '92. ACM, 1992, pp. 287–294. ISBN: 0-89791-497-X. DOI: `10.1145/130385.130417`. URL: `http://doi.acm.org/10.1145/130385.130417`.

[389] Vladimir Shakirov. "Review of state-of-the-arts in artificial intelligence with application to AI safety problem". In: *CoRR* abs/1605.04232 (2016). URL: `https://arxiv.org/pdf/1605.04232v2`.

[390] Shai Shalev-Shwartz. "Online Learning and Online Convex Optimization". In: *Foundations and Trends® in Machine Learning* 4.2 (2012), pp. 107–194. ISSN: 1935-8237. DOI: `10.1561/2200000018`. URL: `http://www.cs.huji.ac.il/~shais/papers/OLsurvey.pdf`.

[391] U. Shalit, F. Johansson, and D. Sontag. "Estimating individual treatment effect: generalization bounds and algorithms". In: *ArXiv e-prints* (June 2016). arXiv: `1606.03976 [stat.ML]`.

[392] Murray Shanahan. *The Technological Singularity.* Forthcoming. MIT Press, 2015.

[393] Jaeho Shin et al. "Incremental Knowledge Base Construction Using DeepDive". In: *Proceedings of the VLDB Endowment* 8.11 (July 2015), pp. 1310–1321. ISSN: 2150-8097. DOI: `10.14778/2809974.2809991`. URL: `http://dx.doi.org/10.14778/2809974.2809991`.

[394] Md Amran Siddiqui et al. "Finite Sample Complexity of Rare Pattern Anomaly Detection". In: *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence, UAI 2016, June 25-29, 2016, New York City, NY, USA.* AUAI Press, 2016, pp. 686–695. URL: `http://auai.org/uai2016/proceedings/papers/226.pdf`.

[395] Md Amran Siddiqui et al. "Sequential Feature Explanations for Anomaly Detection". In: *CoRR* abs/1503.00038 (2015). 9 pages, 4 figures and submitted to KDD 2015. URL: `https://pdfs.semanticscholar.org/cb15/e3855bb420a7a73eadb0c4d38fd1095dc209.pdf`.

[396] Herbert A. Simon. "Rational choice and the structure of the environment". In: *Psychological Review* 63.2 (1956), pp. 129–138. URL: `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.545.5116&rep=rep1&type=pdf`.

[397] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps". In: *CoRR* abs/1312.6034 (2013). URL: `http://www.robots.ox.ac.uk/~vgg/publications/2014/Simonyan14a/simonyan14a.pdf`.

[398] T. A. B. Snijders and R. J. Bosker. *Multilevel Analysis: an Introduction to Basic and Advanced Multilevel Modeling.* 2nd ed. Sage, 2011.

[399] David Snowden. "Multi-ontology sense making: a new simplicity in decision making". In: *Journal of Innovation in Health Informatics* 13.1 (2005), pp. 45–53. URL: `http://hijournal.bcs.org/index.php/jhi/article/viewFile/578/590`.

[400] Nate Soares. *Formalizing Two Problems of Realistic World-Models.* Tech. rep. Machine Intelligence Research Institute, 2015. URL: `https://intelligence.org/files/RealisticWorldModels.pdf`.

[401] Nate Soares. "New Papers Dividing Logical Uncertainty Into Two Subproblems". In: *MIRI* (2016). URL: `https://intelligence.org/2016/04/21/two-new-papers-uniform/`.

[402] Nate Soares. *The Value Learning Problem.* Tech. rep. Paper presented at IJCAI-16 Ethics for AI Workshop, New York, July 9. Machine Intelligence Research Institute, 2016. URL: `https://intelligence.org/files/ValueLearningProblem.pdf`.

[403] Nate Soares and Benja Fallenstein. *Aligning Superintelligence with Human Interests: A Technical Research Agenda.* Tech. rep. Forthcoming 2017 in "The Technological Singularity: Managing the Journey" Jim Miller, Roman Yampolskiy, Stuart J. Armstrong, and Vic Callaghan, Eds. Berkeley,CA: Machine Intelligence Research Institute. Machine Intelligence Research Institute, 2014. URL: `http://intelligence.org/files/TechnicalAgenda.pdf`.

[404] Nate Soares and Benja Fallenstein. *Questions of Reasoning Under Logical Uncertainty.* Tech. rep. Machine Intelligence Research Institute, 2014. URL: http://intelligence.org/files/QuestionsLogicalUncertainty.pdf.

[405] Nate Soares and Benja Fallenstein. "Toward Idealized Decision Theory". In: *CoRR* abs/1507.01986 (2015). URL: https://pdfs.semanticscholar.org/40b3/bbe8d3e0ff66caae3217f4b2fc0e71fd01e2.pdf.

[406] Nate Soares and Benja Fallenstein. "Two Attempts to Formalize Counterpossible Reasoning in Deterministic Settings". In: *In Artificial General Intelligence: 8th International Conference, AGI 2015.* Machine Intelligence Research Institute. Springer International Publishing, 2015, pp. 156–165. URL: https://intelligence.org/files/CounterpossibleReasoning.pdf.

[407] Nate Soares et al. "Corrigibility". In: *AAAI-15 Workshop on AI and Ethics.* 2015. URL: http://aaai.org/ocs/index.php/WS/AAAIW15/paper/viewFile/10124/10136.

[408] Ray J. Solomonoff. "A formal theory of inductive inference. Part I". In: *Information and Control* 7.1 (1964), pp. 1 –22. ISSN: 0019-9958. DOI: http://dx.doi.org/10.1016/S0019-9958(64)90223-2. URL: http://www.sciencedirect.com/science/article/pii/S0019995864902232/pdf?md5=528dcc7a51a90f7254fe06f76ea5f007&pid=1-s2.0-S0019995864902232-main.pdf.

[409] Ray J Solomonoff. "A formal theory of inductive inference. Part II". In: *Information and Control* 7.2 (1964), pp. 224 –254. ISSN: 0019-9958. DOI: http://dx.doi.org/10.1016/S0019-9958(64)90131-7. URL: http://www.sciencedirect.com/science/article/pii/S0019995864901317.

[410] J. Song and J. Alves-Foss. "The DARPA Cyber Grand Challenge: A Competitor's Perspective". In: *IEEE Security Privacy* 13.6 (2015), pp. 72–76. ISSN: 1540-7993. DOI: 10.1109/MSP.2015.132. URL: https://www.researchgate.net/profile/Jim_Alves-Foss/publication/286490027_The_DARPA_cyber_grand_challenge_A_competitor%27s_perspective/links/56778d1908ae502c99d30b3c.

[411] Kaj Sotala. *Concept Learning for Safe Autonomous AI.* Tech. rep. 2015.

[412] Kaj Sotala. *Defining Human Values for Value Learners.* Tech. rep. 2016.

[413] Kaj Sotala and Roman V Yampolskiy. "Responses to catastrophic AGI risk: a survey". In: *Physica Scripta* 90.1 (2015), p. 018001. URL: http://stacks.iop.org/1402-4896/90/i=1/a=018001.

[414] Siddharth Srivastava et al. "First-order Open-universe POMDPs". In: *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence.* UAI'14. Quebec City, Quebec, Canada: AUAI Press, 2914, pp. 742–751. ISBN: 978-0-9749039-1-0. URL: https://people.eecs.berkeley.edu/~russell/papers/uai14-oupomdp.pdf.

[415] Wolfgang Stark. "Ethical Instruments for sustainable development: Principles, Axioms and an Ensemble of Criteria". In: *Symposium Sustainable Development and a New System of Societal Values.* 2001, p. 41. URL: https://www.researchgate.net/profile/Benjamin_Karatzoglou/publication/259011163_Symposium_on_Sustainable_Development_and_a_New_System_of_Societal_Values_Interactions_between_societal_values_and_sustainability_in_the_Greek_tourist_regions/links/0c960529c3fc8873c4000000.pdf#page=41.

[416] Jacob Steinhardt. *Long-Term and Short-Term Challenges to Ensuring the Safety of AI Systems.* Blog post. 2015. URL: https://jsteinhardt.wordpress.com/2015/06/24/long-term-and-short-term-challenges-to-ensuring-the-safety-of-ai-systems/.

[417] Jacob Steinhardt and Percy Liang. "Unsupervised Risk Estimation Using Only Conditional Independence Structure". In: *CoRR* abs/1606.05313 (2016). URL: https://papers.nips.cc/paper/6201-unsupervised-risk-estimation-using-only-conditional-independence-structure.pdf.

[418] Jacob Steinhardt and Percy Liang. "Unsupervised Risk Estimation with only Structural Assumptions". In: (2016). URL: http://cs.stanford.edu/~jsteinhardt/publications/risk-estimation/preprint.pdf.

[419] Jacob Steinhardt, Gregory Valiant, and Moses Charikar. "Avoiding Imposters and Delinquents: Adversarial Crowdsourcing and Peer Prediction". In: *CoRR* abs/1606.05374 (2016). URL: http://cs.stanford.edu/~jsteinhardt/publications/crowdsourcing/paper.pdf.

[420] Bas R. Steunebrink, Kristinn R. Thórisson, and Jürgen Schmidhuber. “Growing Recursive Self-Improvers”. In: *Artificial General Intelligence: 9th International Conference, AGI 2016, New York, NY, USA, July 16-19, 2016, Proceedings*. Ed. by Bas Steunebrink, Pei Wang, and Ben Goertzel. Cham: Springer International Publishing, 2016, pp. 129–139. ISBN: 978-3-319-41649-6. DOI: 10.1007/978-3-319-41649-6_13. URL: http://people.idsia.ch/~steunebrink/Publications/AGI16_growing_recursive_self-improvers.pdf.

[421] Bas Steunebrink. *Experience-Basd AI (EXPAI)*. 2016. URL: https://intelligence.org/files/csrbai/steunebrink-slides.pdf.

[422] Andreas Stuhlmüller, Jacob Taylor, and Noah Goodman. “Learning Stochastic Inverses”. In: *Advances in Neural Information Processing Systems 26*. Ed. by C. J. C. Burges et al. Curran Associates, Inc., 2013, pp. 3048–3056. URL: http://papers.nips.cc/paper/4966-learning-stochastic-inverses.pdf.

[423] Sainbayar Sukhbaatar et al. “End-To-End Memory Networks”. In: *CoRR* abs/1503.08895 (2015). URL: https://arxiv.org/pdf/1503.08895v5.

[424] John P Sullins. “Introduction: Open questions in roboethics”. In: *Philosophy & Technology* 24.3 (2011), pp. 233–238. URL: http://link.springer.com/content/pdf/10.1007%2Fs13347-011-0043-6.pdf.

[425] Ron Sun. *Cognition and Multi-Agent Interaction: From Cognitive Modeling to Social Simulation*. Cambridge University Press, 2005. URL: http://www.cambridge.org/us/academic/subjects/psychology/cognition/cognition-and-multi-agent-interaction-cognitive-modeling-social-simulation?format=HB&isbn=9780521839648.

[426] Ron Sun. *Grounding Social Sciences in Cognitive Sciences*. MIT Press, 2012.

[427] Adith Swaminathan and Thorsten Joachims. “Batch Learning from Logged Bandit Feedback Through Counterfactual Risk Minimization”. In: *The Journal of Machine Learning Research* 16.1 (Jan. 2015), pp. 1731–1755. ISSN: 1532-4435. URL: http://dl.acm.org/citation.cfm?id=2789272.2886805.

[428] Umar Syed and Robert Schapire. “Imitation Learning with a Value-Based Prior”. In: *Proceedings of the Twenty-Third Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-07)*. AUAI Press, 2007, pp. 384–391. URL: http://rob.schapire.net/papers/SyedSchapireUAI2007.pdf.

[429] Christian Szegedy et al. *Intriguing properties of neural networks*. Preprint: arXiv:1312.6199. 2013. URL: https://cs.nyu.edu/~zaremba/docs/understanding.pdf.

[430] István Szita and Csaba Szepesvari. “Agnostic KWIK learning and efficient approximate reinforcement learning.” In: *COLT*. Ed. by Sham M. Kakade and Ulrike von Luxburg. Vol. 19. JMLR Proceedings. JMLR.org, 2011, pp. 739–772. URL: http://dblp.uni-trier.de/db/journals/jmlr/jmlrp19.html#SzitaS11.

[431] Behzad Tabibian et al. “Distilling Information Reliability and Source Trustworthiness from Digital Traces”. In: *CoRR* abs/1610.07472 (2016). URL: https://arxiv.org/pdf/1610.07472v2.

[432] Aviv Tamar, Yonatan Glassner, and Shie Mannor. “Policy Gradients Beyond Expectations: Conditional Value-at-Risk”. In: *CoRR* abs/1404.3862 (2014). URL: http://arxiv.org/abs/1404.3862.

[433] Carmen Tanner and Douglas L. Medin. “Protected values: No omission bias and no framing effects”. In: *Psychonomic Bulletin & Review* 11.1 (2004), pp. 185–191. ISSN: 1531-5320. DOI: 10.3758/BF03206481. URL: http://link.springer.com/content/pdf/10.3758%2FBF03206481.pdf.

[434] Nick Tarleton. *Coherent Extrapolated Volition: A Meta-Level Approach to Machine Ethics*. Tech. rep. Machine Intelligence Research Institute, 2010. URL: https://intelligence.org/files/CEV-MachineEthics.pdf.

[435] Jessica Taylor. *Discussion of bridging consistent and inconsistent preferences*. Discussion. 2016.

[436] Jessica Taylor. “Quantilizers: A Safer Alternative to Maximizers for Limited Optimization”. In: *The Workshops of the Thirtieth AAAI Conference on Artificial Intelligence AI, Ethics, and Society*. Machine Intelligence Research Institute. 2016, pp. 124–131. URL: https://intelligence.org/files/QuantilizersSaferAlternative.pdf.

[437] Jessica Taylor et al. *Alignment for Advanced Machine Learning Systems*. Tech. rep. Machine Intelligence Research Institute, 2016. URL: https://intelligence.org/files/AlignmentMachineLearning.pdf.

[438] Matthew E. Taylor and Peter Stone. "Transfer Learning for Reinforcement Learning Domains: A Survey". In: *J. Mach. Learn. Res.* 10 (Dec. 2009), pp. 1633–1685. ISSN: 1532-4435. URL: http://www.jmlr.org/papers/volume10/taylor09a/taylor09a.pdf.

[439] Max Tegmark. "Friendly Artificial Intelligence: the Physics Challenge". In: *AAAI-15 Workshop on AI and Ethics*. 2015. URL: http://www.aaai.org/ocs/index.php/WS/AAAIW15/paper/download/10149/10138.

[440] Joshua B Tenenbaum et al. "How to grow a mind: Statistics, structure, and abstraction". In: *Science* 331.6022 (2011), pp. 1279–1285. URL: https://web.stanford.edu/~ngoodman/papers/tkgg-science11-reprint.pdf.

[441] Moshe Tennenholtz. "Program equilibrium". In: *Games and Economic Behavior* 49.2 (2004), pp. 363–373. URL: https://ie.technion.ac.il/~moshet/progeqnote4.pdf.

[442] Philip S. Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. "High Confidence Off-policy Evaluation". In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. AAAI'15. AAAI Press, 2015, pp. 3000–3006. ISBN: 0-262-51129-0. URL: http://psthomas.com/papers/Thomas2015.pdf.

[443] Andrea L Thomaz and Cynthia Breazeal. "Transparency and socially guided machine learning". In: *5th Intl. Conf. on Development and Learning (ICDL)*. 2006. URL: http://www.cc.gatech.edu/~athomaz/papers/ThomazBreazeal-ICDL06.pdf.

[444] Adrian Thompson. "Artificial Evolution in the Physical World". In: *In Evolutionary Robotics: From Intelligent Robots to Artificial Life (ER'97*. AAI Books, 1997, pp. 101–125. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.26.6187&rep=rep1&type=pdf.

[445] Kristinn R. Thórisson et al. "About Understanding". In: *Artificial General Intelligence: 9th International Conference, AGI 2016, New York, NY, USA, July 16-19, 2016, Proceedings*. Ed. by Bas Steunebrink, Pei Wang, and Ben Goertzel. Cham: Springer International Publishing, 2016, pp. 106–117. ISBN: 978-3-319-41649-6. DOI: 10.1007/978-3-319-41649-6_11. URL: http://dx.doi.org/10.1007/978-3-319-41649-6_11.

[446] Kristinn R. Thórisson et al. "Why Artificial Intelligence Needs a Task Theory". In: *Artificial General Intelligence: 9th International Conference, AGI 2016, New York, NY, USA, July 16-19, 2016, Proceedings*. Ed. by Bas Steunebrink, Pei Wang, and Ben Goertzel. Cham: Springer International Publishing, 2016, pp. 118–128. ISBN: 978-3-319-41649-6. DOI: 10.1007/978-3-319-41649-6_12. URL: http://dx.doi.org/10.1007/978-3-319-41649-6_12.

[447] Aristide C. Y. Tossou and Christos Dimitrakakis. "Algorithms for Differentially Private Multi-Armed Bandits". In: *13th International Conference on Artificial Intelligence (AAAI 2016)*. AAAI Press, 2016, pp. 2087–2093. URL: http://www.aaai.org/Library/AAAI/aaai16contents.php.

[448] M. Tulio Ribeiro, S. Singh, and C. Guestrin. "Model-Agnostic Interpretability of Machine Learning". In: *ArXiv e-prints* (June 2016). arXiv: 1606.05386 [stat.ML]. URL: https://pdfs.semanticscholar.org/2f99/1be8d35e4c1a45bfb0d646673b1ef5239a1f.pdf.

[449] Matteo Turchetta, Felix Berkenkamp, and Andreas Krause. "Safe exploration in finite Markov Decision Processes with Gaussian processes". In: *Proc. of the Conference on Neural Information Processing Systems (NIPS)*. (to appear). 2016. URL: http://arxiv.org/abs/1606.04753.

[450] R. Turner. "A Model Explanation System: Latest Updates and Extensions". In: *ArXiv e-prints* (2016). arXiv: 1606.09517 [stat.ML]. URL: https://arxiv.org/pdf/1606.09517.pdf.

[451] Tomer Ullman et al. "Help or Hinder: Bayesian Models of Social Goal Inference". In: *Advances in Neural Information Processing Systems 22*. Ed. by Y. Bengio et al. Curran Associates, Inc., 2009, pp. 1874–1882. URL: http://papers.nips.cc/paper/3747-help-or-hinder-bayesian-models-of-social-goal-inference.pdf.

[452] R. Vedantam et al. "Learning Common Sense through Visual Abstraction". In: *2015 IEEE International Conference on Computer Vision (ICCV)*. 2015, pp. 2542–2550. DOI: 10.1109/ICCV.2015.292. URL: http://www.cv-foundation.org/openaccess/content_iccv_2015/papers/Vedantam_Learning_Common_Sense_ICCV_2015_paper.pdf.

[453] R. Velik and D. Bruckner. “Neuro-symbolic networks: introduction to a new information processing principle”. In: *2008 6th IEEE International Conference on Industrial Informatics*. 2008, pp. 1042–1047. DOI: 10.1109/INDIN.2008.4618256. URL: https://publik.tuwien.ac.at/files/PubDat_166316.pdf.

[454] Alfredo Vellido, Josh D Martin-Guerrero, and Paulo J.G. Lisboa. “Making machine learning models interpretable”. In: *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*. ESANN, 2012, pp. 163–172. URL: https://pdfs.semanticscholar.org/ce0b/8b6fca7dc089548cc2e9aaac3bae82bb19da.pdf.

[455] Joel Veness et al. “Reinforcement Learning via AIXI Approximation”. In: *CoRR* abs/1007.2049 (2010). URL: http://arxiv.org/abs/1007.2049.

[456] Vernor Vinge. “The coming technological singularity”. In: *VISION-21 Symposium, NASA Lewis Research Center and the Ohio Aerospace Institute*. NASA CP-10129. 1993. URL: http://www-rohan.sdsu.edu/faculty/vinge/misc/singularity.html.

[457] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer-Verlag New York, Inc., 2005. ISBN: 0387001522.

[458] Wendell Wallach and Colin Allen. *Moral machines: Teaching robots right from wrong*. Oxford University Press, 2008.

[459] Wendell Wallach, Stan Franklin, and Colin Allen. “A Conceptual and Computational Model of Moral Decision Making in Human and Artificial Agents”. In: *Topics in Cognitive Science* 2.3 (July 2010), pp. 454–485. ISSN: 1756-8765. DOI: 10.1111/j.1756-8765.2010.01095.x. URL: http://onlinelibrary.wiley.com/doi/10.1111/j.1756-8765.2010.01095.x/epdf.

[460] Nik Weaver. *Paradoxes of rational agency and formal systems that verify their own soundness*. Preprint. 2013. URL: http://arxiv.org/pdf/1312.3626.pdf.

[461] Philippe Weber et al. “Overview on Bayesian Networks Applications for Dependability, Risk Analysis and Maintenance Areas”. In: *Engineering Applications of Artificial Intelligence* 25.4 (June 2012), pp. 671–682. ISSN: 0952-1976. DOI: 10.1016/j.engappai.2010.06.002. URL: http://dx.doi.org/10.1016/j.engappai.2010.06.002.

[462] Daniel Weld and Oren Etzioni. “The first law of robotics (a call to arms)”. In: *AAAI*. Vol. 94. 1994, pp. 1042–1047. URL: http://homes.cs.washington.edu/~weld/papers/first-law-aaai94.pdf.

[463] Keenon Werling et al. “On-the-Job Learning with Bayesian Decision Theory”. In: *Advances in Neural Information Processing Systems 28*. Ed. by C. Cortes et al. Curran Associates, Inc., 2015, pp. 3447–3455. URL: http://papers.nips.cc/paper/5860-on-the-job-learning-with-bayesian-decision-theory.pdf.

[464] Wolfram Wiesemann, Daniel Kuhn, and Rustem. “Robust Markov decision processes”. In: *Mathematics of Operations Research* 38.1 (2013), pp. 153–183.

[465] Alan FT Winfield, Christian Blum, and Wenguo Liu. “Towards an Ethical Robot: Internal Models, Consequences and Ethical Action Selection”. In: *Advances in Autonomous Robotics Systems*. Springer, 2014, pp. 85–96. URL: https://pdfs.semanticscholar.org/5608/b9ed2454a8788d0ecaab9a10d87b78125d
pdf.

[466] Terry Winograd. “Architectures for Context”. In: *Hum.-Comput. Interact.* 16.2 (Dec. 2001), pp. 401–419. ISSN: 0737-0024. DOI: 10.1207/S15327051HCI16234_18. URL: https://pdfs.semanticscholar.org/d275/94ddf7e7849a87f940b25db9849d668a93b3.pdf.

[467] Laurenz Wiskott and Terrence J. Sejnowski. “Slow feature analysis: unsupervised learning of invariances”. In: *Neural Comput* 14.4 (Apr. 2002), pp. 715–770. DOI: 10.1162/089976602317318938. URL: http://www.ncbi.nlm.nih.gov/pubmed/11936959.

[468] AD Wissner-Gross and CE Freer. “Causal entropic forces”. In: *Physical review letters* 110.16 (2013), p. 168702. URL: http://math.mit.edu/~freer/papers/PhysRevLett_110-168702.pdf.

[469] Zhen Xu et al. “Incorporating Loose-Structured Knowledge into LSTM with Recall Gate for Conversation Modeling”. In: *CoRR* abs/1605.05110 (2016). URL: https://arxiv.org/pdf/1605.05110v1.

[470] R. V. Yampolskiy. “Verifier Theory and Unverifiability”. In: *ArXiv e-prints* (Sept. 2016). arXiv: 1609.00331 [cs.AI]. URL: https://arxiv.org/pdf/1609.00331v3.

[471] Roman V. Yampolskiy. “Artificial Intelligence Safety and Cybersecurity: a Timeline of AI Failures”. In: *CoRR* abs/1610.07997 (2016). URL: http://arxiv.org/abs/1610.07997.

[472] Roman V. Yampolskiy. “From Seed AI to Technological Singularity via Recursively Self-Improving Software”. In: *CoRR* abs/1502.06512 (2015). URL: http://arxiv.org/abs/1502.06512.

[473] Roman V. Yampolskiy. “Taxonomy of Pathways to Dangerous Artificial Intelligence”. In: *AAAI Workshop: AI, Ethics, and Society*. Vol. WS-16-02. AAAI Workshops. AAAI Press, 2016. URL: http://www.aaai.org/ocs/index.php/WS/AAAIW16/paper/download/12566/12356.

[474] Roman V. Yampolskiy. “Utility function security in artificially intelligent agents”. In: *Journal of Experimental & Theoretical Artificial Intelligence* 26.3 (2014), pp. 373–389. DOI: 10.1080/0952813X.2014.895114. eprint: http://dx.doi.org/10.1080/0952813X.2014.895114. URL: http://dx.doi.org/10.1080/0952813X.2014.895114.

[475] Roman V. Yampolskiy and Joshua Fox. “Artificial General Intelligence and the Human Mental Model”. In: *Singularity Hypotheses: A Scientific and Philosophical Assessment*. Ed. by Amnon H. Eden et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 129–145. ISBN: 978-3-642-32560-1. DOI: 10.1007/978-3-642-32560-1_7. URL: https://intelligence.org/files/AGI-HMM.pdf.

[476] Roman Yampolskiy. “Leakproofing the Singularity: Artificial Intelligence Confinement Problem”. In: *Journal of Consciousness Studies* 19.1-2 (2012), pp. 1–2. URL: http://cecs.louisville.edu/ry/LeakproofingtheSingularity.pdf.

[477] Yezhou Yang et al. “Robot Learning Manipulation Action Plans by ”Watching” Unconstrained Videos from the World Wide Web”. In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. AAAI’15. Austin, Texas: AAAI Press, 2015, pp. 3686–3692. ISBN: 0-262-51129-0. URL: https://www.umiacs.umd.edu/~yzyang/paper/YouCookMani_CameraReady.pdf.

[478] Eliezar Yudkowsky. “Artificial Intelligence as a Positive and Negative Factor in Global Risk”. In: *Global Catastrophic Risks* 1 (2008), p. 303. URL: https://intelligence.org/files/AIPosNegFactor.pdf.

[479] Eliezer Yudkowsky. *Coherent Extrapolated Volition*. Tech. rep. Machine Intelligence Research Institute, 2004. URL: https://intelligence.org/files/CEV.pdf.

[480] Eliezer Yudkowsky. “Complex Value Systems in Friendly AI”. In: *Artificial General Intelligence: 4th International Conference, AGI 2011, Mountain View, CA, USA, August 3-6, 2011. Proceedings*. Ed. by Jürgen Schmidhuber, Kristinn R. Thórisson, and Moshe Looks. Springer Berlin Heidelberg, 2011, pp. 388–393. ISBN: 978-3-642-22887-2. DOI: 10.1007/978-3-642-22887-2_48. URL: http://dx.doi.org/10.1007/978-3-642-22887-2_48.

[481] Eliezer Yudkowsky and Marcello Herreshoff. “Tiling Agents for Self-Modifying AI, and the Löbian Obstacle”. 2013. URL: http://intelligence.org/files/TilingAgents.pdf.

[482] Fouad Zablith. “Ontology evolution: a practical approach”. In: *Workshop on Matching and Meaning at Artificial Intelligence and Simulation of Behaviour*. 2009.

[483] Tom Zahavy, Nir Ben Zrihem, and Shie Mannor. “Graying the Black Box: Understanding DQNs”. In: *Proceedings of the 33rd International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: WCP volume 48*. JMLR.org, 2016, pp. 1899–1908. URL: http://jmlr.org/proceedings/papers/v48/zahavy16.pdf.

[484] Matthew D. Zeiler and Rob Fergus. “Visualizing and Understanding Convolutional Networks”. In: *Computer Vision – ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*. Ed. by David Fleet et al. Springer International Publishing, 2014, pp. 818–833. ISBN: 978-3-319-10590-1. DOI: 10.1007/978-3-319-10590-1_53. URL: http://www.cs.nyu.edu/~fergus/papers/zeilerECCV2014.pdf.

[485] H. Zhang et al. “StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks”. In: *ArXiv e-prints* (Dec. 2016). arXiv: 1612.03242 [cs.CV]. URL: https://arxiv.org/pdf/1612.03242v1.

[486] S. Zhang, M. Cao, and M. K. Camlibel. “Upper and Lower Bounds for Controllable Subspaces of Networks of Diffusively Coupled Agents”. In: *IEEE Transactions on Automatic Control* 59.3 (2014), pp. 745–750. ISSN: 0018-9286. DOI: 10.1109/TAC.2013.2275666. URL: https://pdfs.semanticscholar.org/530a/1373abd714070cb9b65ab6534216b98512f5.pdf.

[487] Yuchen Zhang et al. “Spectral Methods meet EM: A Provably Optimal Algorithm for Crowdsourcing”. In: *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*. 2014, pp. 1260–1268. URL: http://papers.nips.cc/paper/5431-spectral-methods-meet-em-a-provably-optimal-algorithm-for-crowdsourcing.

[488] Zuhe Zhang, Benjamin I. P. Rubinstein, and Christos Dimitrakakis. “On the Differential Privacy of Bayesian Inference”. In: *Thirtieth International Conference on Artificial Intelligence (AAAI 2016)*. 2016. URL: http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/viewFile/12129/11885.

[489] Brian D. Ziebart et al. “Maximum Entropy Inverse Reinforcement Learning”. In: *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence (2008)*. 2008, pp. 1433–1438. URL: http://www.aaai.org/Papers/AAAI/2008/AAAI08-227.pdf.