NSF Convergence Accelerator Workshop Report

Inaugural Workshop on Provably Safe and Beneficial AI (PSBAI)

Funded by NSF Grant #2230996

PI: Russell, Stuart J, University of California, Berkeley

January 25, 2023

Motivation and Overview

The use of increasingly powerful AI systems for a broad range of purposes carries with it significant risks. These issues have been discussed from a variety of perspectives including technology, law, economics, medicine, sociology, politics, and ethics, and under a variety of names that indicate more specific concerns, such AI safety, AI ethics, responsible AI, beneficial AI, human-centered AI, AI alignment, and so on. No matter which domain one considers, there are at least two profound challenges to the safe and beneficial use of AI technology: deciding how AI systems should behave, and building them to conform to these requirements. This document proposes a multi-disciplinary, cross-sector research program to address the technical challenges of AI safety at a fundamental level.

The goal of this research program is to develop the technology to enable the creation of highly capable AI systems that provably do what they are designed to do, while handling the inevitable gaps in the specifications we have given them in a safe and beneficial way.

The Inaugural Workshop on Provably Safe and Beneficial AI (PSBAI) was held in person in Berkeley, California on the weekend of October 7-9, 2022, with 51 primary attendees from a range of disciplines. Its sister workshop on Ethical Design of AIs (EDAI) was held virtually with a plenary session on September 22 and three subsequent working group sessions. These two workshops covered complementary aspects of the challenge of ensuring that AI is safe and beneficial when integrated into individual lives and social systems. EDAI included sophisticated consideration of ethical principles, human-centered design, and AI governance, as well as the articulation of many domain-specific challenges. PSBAI attacked the corresponding challenge of making it possible to build systems that can implement the outputs of these design processes. **In short, no AI policy will be successful without the technical capability to implement it.**

This is much more than just a technical problem. As the EDAI workshop report noted, “many AI systems use radically different concepts than humans, and so our ethical theories cannot necessarily be translated into AI algorithms.” In other words, it may not be possible to develop AI technology that can implement existing ethical principles on their own terms, due to a stark disconnect between policy and technical vocabularies. Instead, we need to discover what kinds of technical building blocks can be used to create systems with useful safety guarantees, and orient policy and design approaches around them. **While AI governance and ethics research will define what systems are desirable, AI safety research will create the vocabulary to express normative ideas in implementable forms.**

This is a particularly imminent challenge because current AI technologies are not founded on safe principles. Deep learning models are not *designed* in the traditional sense. Because a trained deep neural network performs its task according to unknown principles, it is very hard to ensure safety in all situations. And in practice, every kind of deep learning system shows severe vulnerabilities and unpredictable failure modes—consider, for example, Carter et al. (2021), Gleave et al. (2020), and the many thousands of vulnerabilities already revealed by ChatGPT. Even the much-vaunted "superhuman" Go programs turn out to lose consistently (and for unknown reasons) to almost child-like strategies that any human expert can defeat easily (Wang et al., 2022).

While deep learning systems often fail to meet specifications, they, as well as more traditional system designs in all areas of engineering, are subject to another kind of failure: *mis-specification of objectives*. Both theoretically and in practice, this leads to the potential for catastrophic misalignment between true human preferences and the AI system's objectives and behavior—even for a fully "verified" system. The optimization of clickthrough or engagement metrics by social media algorithms, with potentially disastrous consequences for human society (Stray et. al. 2022), serves as an early warning; furthermore, *improving* the optimization of incorrect objectives leads to *worse* outcomes (Zhuang and Hadfield-Menell, 2020).

Addressing these issues will require a concerted effort from multiple disciplines, ranging from mathematically rigorous branches of computer science, engineering, and statistics to conceptually challenging areas of moral philosophy and behavioral economics. In Section 1, we describe the major threads of work in this domain, drawing out the connections between disciplines and sectors. Section 2 analyzes the research themes from the workshop and Section 3 crystallizes these themes into an overall research program with a set of five interrelated research areas. Appendix A gives a detailed report of the session presentations from which this report was synthesized.

1. Technical Foundations for an Interdisciplinary Problem

The problem of making AI safe and beneficial spans almost every human discipline—not just because AI is so widely applied, but also because the problem itself has so many facets. Computer science, statistics, control theory, and safety engineering provide the scientific and technical underpinnings. Behavioral economics, mechanism design, and game theory are foundational tools for reasoning about the actions of automata and people together. Psychology and sociology clarify human needs and help us understand what happens when people and societies interact with AI. Human-centered design, value-sensitive design, and participatory design supply methodologies for considering the multitude of human voices and needs. Law and policy consider how AI systems might integrate with, and be guided by, legal and governance frameworks. Philosophy, political science, and ethics directly address the high-level question of "what is good for a machine to do?" In addition to all of these general fields, specific AI applications demand deep domain knowledge in medicine, journalism, defense, public policy, etc.

This problem is also multi-sector. Although PSBAI was organized by an academic institution, scholars on their own cannot redirect the commercial development of AI. Industry will need to draw on the research produced in University labs, and many of the most challenging problems cannot be addressed without deep industry-academic research collaborations. The direction of research cannot be set in a vacuum, but must be guided by civil society organizations that represent those potentially affected by AI systems. Finally, government and policy actors must set the regulatory standards that protect consumers, ensure security, and promote a flourishing and competitive AI industry.

The PSBAI workshop was an opportunity for focussed work on the technical end of these problems. The 51 attendees, many of them leading experts in their fields, spanned a range of technical disciplines including artificial intelligence, programming languages, formal methods, control theory, game theory, statistics, and safety engineering, but also included key interdisciplinary researchers in philosophy, health care, defense, media, and law. There were several senior industry researchers in attendance, as well as government representatives. Attendees came from the UK, Canada, and Australia, as well as the US. The full attendee list is given in Appendix B.

The workshop focused on three closely interrelated technical themes:

- General AI safety: methods for ensuring that AI systems are safe and beneficial for humans, regardless of how capable they become.
- Well-founded AI system design: building AI systems from semantically well-defined components with rigorous compositional properties, with a particular focus on probabilistic programming as an enabling technology.
- Formal methods for verification and synthesis: methods providing rigorous guarantees of correctness for software instantiations of well-founded agent designs.

Overview talks were given on each of these themes by Stuart Russell (UC Berkeley), Vikash Mansinghka (MIT), and Sanjit Seshia (UC Berkeley), respectively. In addition, because policy and regulation are necessary adjuncts of any technical approach to safety, Gillian Hadfield (Toronto) gave an overview of the current AI regulatory landscape and future possibilities. Summaries of these talks can be found in Appendix A.

2. Emerging Research Directions

The talks and discussions revealed a rich and highly interconnected set of research questions that will require contributions from a wide range of disciplines. One foundational theme was the need to remove the standard assumption that the objective is fixed and perfectly known by the system. This assumption underlies almost all existing AI technologies. Previous theoretical work (Hadfield-Menell et. al., 2016, 2017; Russell, 2019) suggests that explicitly representing uncertainty over human preferences leads naturally to a variety of safe and beneficial AI behavior, including asking the human for clarification or permission. This has become known as the “assistance game” paradigm. Solving an assistance game with a human not only involves learning what the human prefers, but also learning to resolve or avoid coordination problems, e.g., deciding and signaling who is responsible for what, and when, and in what way.

Because many basic technologies in AI—such as search, planning, and reinforcement learning—assume a fixed, perfectly known objective, they will need to be redeveloped on a new, broader foundation.

Although many at the workshop had the belief that several kinds of benevolent behavior, such as asking permission and allowing oneself to be switched off, can only be exhibited in a natural sense (i.e., without pre-scripting) by agents that are uncertain about human objectives, we need a more precise explication of this view. Also, some suggested it might be possible to develop a robust (e.g., maximin) solution concept for assistance games and to reduce the need for fully specified priors over human preferences.

Participants noted the daunting prospect of creating realistic models of actual human preferences—a topic that would benefit from both conceptual development and experimental research. Challenging problems include the complexity of modeling real human preferences which can be incomplete, inconsistent, dynamic, and highly contextual, and the related problem of inferring preferences from human behavior, natural language, video, etc. AI systems that serve multiple users will also need a philosophically grounded and empirically robust theory of social aggregation, in order to be able to decide what to do when people inevitably disagree. This interdisciplinary effort will require the involvement of cognitive scientists, behavioral economists, sociologists, ethnographers, political scientists, etc. Some also pointed to the possibility of using PSBAI concepts to develop a broad theoretical foundation for human-computer interaction, including the elucidation of semantics for requests, commands, prohibitions, and choices.

Aside from assistance games, there are a plethora of other ideas for achieving alignment being pursued by a highly creative research community (Irving et al., 2018; Christiano et al., 2018; Hendrycks et al., 2021; Hubinger, 2020). These generally draw on the idea that, from the point of view of computational complexity, checking is easier than generating. This means that less capable AI systems (and perhaps humans) can check the outputs of more powerful AIs for safety or correctness, leading to the possibility of an iterated development process that maintains safety at each stage.

Well-founded AI is an approach that seeks to build systems from semantically well-defined, rigorously composed elements. Several promising directions emerged from the workshop. Probabilistic programming languages (PPL) already show superior performance with less training data for certain computer vision problems (Gothoskar et al., 2021) and have already been deployed successfully for global nuclear monitoring (Arora et al., 2013). This suggests the need for a full-scale engineering and experimental effort to develop robust, usable PPL platforms and systems and explore their applicability. Potential applications for PPL-derived safe AI include tracking many coupled objects from video input (as in self-driving vehicles), cataloging and analyzing human behaviors and activities (as would be necessary for common-sense interaction), natural language processing, computational systems biology, computer security, and intelligence analysis. Extending PPLs from inference systems into full-blown agents is also a largely open problem, one that preliminary work suggests is challenging (Srivastava et al., 2014).

Several participants emphasized the need to develop a robust safety methodology for not-yet-well-founded AI technologies such as deep convolutional nets and large language models (LLMs), as these are likely to continue to be used in real applications until better well-founded methods become available. For example, can we tell what LLMs believe and whether it is true? Can we get them to believe true facts and reject false ones? How can we develop scalable human oversight for LLMs and measure its effectiveness? Can we ground LLMs by connecting them to real or simulated environments such that linguistic and sensory inputs can co-refer?

One of the most exciting aspects of the workshop was the degree of convergence between formal methods and other areas. There is overlap with control theory research on the topic of safety, with the application of formal methods to adaptive systems (for example in proving robustness of suitably modified deep learning systems to small input perturbations), and with the use of formal program verification methods to prove a PPL inference system correct. These synergies bode well for the goal of being able to design verified, well-founded AI systems—undoubtedly restricted in functionality at first, but providing the basis for ongoing development while maintaining safety.

If there is a master strategy to developing all of these approaches, it is: start with pieces that are small, narrow in scope, and safe, then build up to systems that are big, broad in scope, and safe. Scaling up towards big and safe systems may be able to call upon assume-guarantee methods for verifying complex systems composed from smaller verified subsystems (Kwiatkowska et al. 2010). In AI, composite systems have long been used—for example, in robotics with components for perception, obstacle avoidance, navigation, task planning, low-level control, health monitoring, and so on. There is as yet, however, no theory of agent architecture: it's "my boxes and arrows are better than your boxes and arrows." The concept of bounded optimality (Russell & Subramanian, 1995; Gershman et al., 2015) offers a way forward and some simple examples of optimal composition have already been obtained; roughly speaking, one proves that some configuration of program architecture A outperforms (in an asymptotic sense) all configurations of architecture B. It may also be possible to use mechanism design ideas to build human-aligned systems from not-necessarily-aligned components.

As the provable properties become stronger, that will also enable stronger regulation on the design, testing, and verification of systems—for example, systems without the required certificates may become uninsurable or illegal in certain contexts. In the long run, the necessary complement of provably safe and beneficial AI is a means of preventing unsafe and harmful AI from being deployed. This becomes increasingly important as AI systems become more capable. Gillian Hadfield's discussion of regulation touched on ideas of certification and "AI passports" that would prevent non-certified AI systems from gaining access to hardware and network resources.[1] This would undoubtedly open up many new avenues for research in computer and network architecture and cybersecurity as well as formal methods, and spur the development of both hard- and soft-law regulatory approaches.

There was broad agreement that further progress in all of these areas would benefit from developing prototype systems for more realistic scenarios—e.g., learning to assist other agents in increasingly complex simulated environments. Near-term applications might include robotics, personal digital assistants, and beneficial recommendations in social media. Recommenders are a particularly challenging environment because platforms are populated by many human and machine agents with preferences that may conflict. Convincing prototypes would also help dispel the perception that safe AI is always going to lag behind commercial AI systems developed within the standard (fixed, known objective) model. Rather, safe AI, like safe passenger aircraft and safe nuclear power stations, will become a commercial necessity.

[1] Some participants pointed to the "Trusted Computing Platform Alliance" (TCPA) and Microsoft's Palladium initiative, both active in the early 2000s, as examples of efforts along these lines that might be ripe for revival, given the enormous increase in cybercrime and cyberwarfare since that period.

3. Recommendations for a PSBAI Research Track

As noted, this is a convergent research topic that spans sectors, as scholars on their own cannot redirect the development of AI: industry practitioners will need to use the research produced in labs, and many of the most challenging problems cannot be addressed without deep industry-academic research collaborations. The direction of research must be guided by civil society organizations who are in direct contact with and represent those potentially affected by AI systems. Finally, government and policy actors must set the regulatory standards that protect consumers, ensure security, and promote a flourishing and competitive AI industry.

A research program leading eventually to a secure digital ecosystem of well-founded and provably safe and beneficial AI systems would combine the efforts of partnerships among several disciplines, focusing on five interrelated areas:

1. *Approaches to beneficial AI*: This area will focus on ways to ensure that AI systems' behavior is actually beneficial, particularly when exact specifications of what is "beneficial" are not available. For example, in the emerging paradigm of "assistance games" the AI system is designed to be initially unsure what the human wants, which leads naturally to behavior such as asking for clarification, deferring to human judgment, etc. Other research paradigms addressing the same questions are certainly possible, such as iterative improvement strategies where weaker systems check the outputs of stronger systems, or the use of mechanism design ideas—incentive strategies—to build human-aligned systems from not-necessarily-aligned components.

Core questions include: how to represent partial and uncertain knowledge of human preference structures; how to define beneficial when AI systems act on behalf of multiple humans whose own preferences are uncertain, dynamic, and potentially opposed; how to interpret human behavior and feedback in terms of underlying preference structures; how to cooperate successfully with humans while learning about their preferences; how to extend current AI technologies (such as problem-solving search, planning, reinforcement learning, and synthesizing safe controllers) to allow for uncertainty over preferences and the interactive flow of information about preferences from humans to machines; how to ensure that multiple machine agents cooperate successfully in helping humans; beneficial-AI frameworks for important application areas such as recommender systems and interactive natural-language systems; and scalable algorithms for all the preceding tasks.

2. *Well-founded AI system designs*: Safety and other related properties are easier to assure when AI systems are designed from well-understood components—particularly components with well-defined, transparent semantics—connected in well-understood ways. This track seeks to create foundational methods for building AI components (such as inference engines and learned models) that are safe by design.

Formal, symbolic representation languages based on logic or probability theory are obvious technical foundations for such components, as they support reasoning and the accumulation of knowledge. One promising class of such languages is the broad family of probabilistic programming languages (PPLs), which provide universal formalisms for defining complex probability models through code, and mechanisms for inference and learning in those models. Hybrids with "black-box" AI methods such as deep neural networks are also possible, although these may come with weaker guarantees of correctness. Research in this area may include developing robust, efficient, and usable representation and inference platforms; extending these platforms from inference to decision-making agent architectures; demonstrating scalable, high-performance solutions for tasks such as tracking many coupled objects from video input (as in self-driving vehicles); keeping track of and cooperating with human behaviors and activities; navigating the common-sense physical world; natural language processing; scientific applications (for example, computational systems biology); computer security; and intelligence analysis. An important complementary research theme is the development of a comparative theory of agent architectures, capable of supporting claims to the effect that one agent architecture is, under certain conditions, superior to another in terms of efficient use of available data (sample complexity), decision quality, etc.

3. *Formal methods for AI system assurance*: This research track aims to develop methods for proving that an AI system satisfies desired safety properties. While track 2 (well-founded AI) addresses the problem of building components on solid foundations, this track focuses on the problem of specifying an AI system formally, then proving that the system meets those specifications – and that it will do something reasonable even when those specifications are incomplete, as they inevitably are.

Specific research topics may include assume-guarantee methods for verifying complex systems composed from smaller verified subsystems; verification of embedded systems that interact with humans to achieve unspecified human objectives; verification of composite systems, some of whose components may involve statistical learning; flexible verification of customizable and adaptive inference, learning, and decision-making systems. Also of interest would be approaches to ensuring the accuracy or safety of not-yet-well-founded AI technologies based on neural networks, such as large language models (LLMs), possibly through grounding in a subset of guaranteed-true assertions or in a connection to an underlying real or simulated environment. In addition, advances in control theory will enable formal synthesis of provably safe human-interactive controllers.

4. *Provably safe and beneficial AI system prototypes*: Research in this area will integrate ideas from the first three areas to develop and demonstrate fully verified prototypes of provably safe and beneficial AI systems.

It is anticipated that this work will start with simple systems that interact with simple agents and environments e.g., simple goal-seeking agents in simulated environments. As the core technologies develop further, more sophisticated prototypes will become possible, including agents acting in increasingly high-fidelity simulations incorporating generative models of human behavior, systems that work in a lab setting with a small number of real humans, and eventually systems that operate in large online environments such as recommender systems in social media platforms.

5. *Policy, regulation and mechanisms for a secure digital ecosystem*: As provably safe and beneficial AI systems become possible and eventually widely useful, while the underlying AI technologies become more powerful, public policy and regulation will need to keep up

Research in this area will explore the development of regulatory approaches (including guidelines, best practices, and regulatory standards, whether generic or domain-specific) that will encourage or require the use of provably safe and beneficial AI system designs where appropriate. Furthermore, it will become increasingly important to ensure that unverified, unsafe, and possibly malicious AI systems cannot be deployed. Research in this area will explore ideas for software and/or hardware infrastructure that will prevent non-certified AI systems from running in safety critical settings, including limiting access to hardware, network, and data resources. Possible approaches include authority-based certification yielding “AI passports” and decentralized approaches such as proof-carrying code. Research topics may include software self-certification, distribution network architectures, cybersecurity, audit methodologies, and hard- and soft-law to enforce and incentivize all of the above.

We imagine that typical research teams would combine two or more of these areas, typically with core strength in at least one of the first three areas.

4. Summary and Conclusion

The workshop’s high-caliber attendees shared recent work and discussed a wide range of ideas on the convergence topic of provably safe and beneficial artificial intelligence, in the end yielding a compelling family of research programs attentive to many aspects of human-AI interaction, as outlined in the preceding sections.

Based on the material and discussion in the PSBAI workshop, the organizers believe it is now plausible, through cross-discipline collaboration, to create an end-to-end demonstration system: applying methods, languages and tools to design and verify certain types of well-founded intelligent agents that are provably safe, and to deploy them securely. This is now a reasonable three-year goal, given sufficient attention and resources.

Appendix A: Session Reports

The workshop was conducted over 3 days at the University of California Berkeley campus, beginning at noon on October 7, 2022, and ending at 1pm on October 9. There were 51 primary attendees as well as administrative staff, PhD students who acted as note-takers, and Berkeley faculty who joined for small portions of the meeting. The attendees, many of them leading experts in their fields, spanned a range of disciplines (artificial intelligence, programming languages, formal methods, control theory, game theory, philosophy, statistics, safety engineering, law) and career stages. Attendees came from the UK, Canada, and Australia, as well as the US, and there were several government representatives. The gender distribution was more balanced than is typical for computer science and engineering events. The full attendee list is given in Appendix B.

There were four hour-long overview talks, three breakout sessions of approximately two hours each with three subgroups, and 17 lightning talks of 15 minutes each. In addition, active and lively informal discussions took place during breaks and meals. The complete schedule is given in Appendix C.

Overview talks

General AI Safety (Stuart Russell, UC Berkeley)

The talk covered some of the main ideas for ensuring that AI systems are safe and beneficial, even as they become far more capable. Russell observed a crucial distinction between classical, knowledge-based AI system designs and those that result from deep-learning processes (e.g., large language models): as the latter are not *designed* in any meaningful sense and operate according to unknown principles, it is very hard to ensure safety. (And in practice, every kind of deep learning system shows severe vulnerabilities and unpredictable failure modes—see, e.g., Carter et al., 2021; Gleave et al., 2020.) On the other hand, traditional proofs of safety for classically designed systems assume that the specification is correct, and, in AI systems, the specification is translated into the system's internal objective; this leads to the potential for catastrophic misalignment between true human preferences and the AI system's objectives and behavior—even for a fully verified system. The optimization of clickthrough or engagement metrics by social media algorithms, with arguably disastrous consequences for human society, serves as an early warning; furthermore, *improving* the optimization of incorrect objectives leads to *worse* outcomes for humans (Zhuang and Hadfield-Menell, 2020).

A solution may be found in a new class of AI systems that are explicitly uncertain about the true human objectives they should optimize. Russell defined these systems as solving *assistance games*, in which M humans with payoffs $U_1,\dots,U_M$ are assisted by N robots whose payoffs are identical to the aggregate human welfare (say, $\sum_i U_i$, but other forms are possible), but who are a priori *uncertain* about $U_1,\dots,U_M$. Such systems may admit proofs of safe and beneficial behavior without the assumption that human objectives can be completely and correctly explicated (Hadfield-Menell et al., 2016; Russell, 2019). A simple proof was given that agents that solve assistance games generally allow themselves to be switched off by sufficiently rational humans (Hadfield-Menell et al., 2017). Because most current AI methods (including problem-solving search, game-playing, planning, dynamic programming, reinforcement learning, adaptive control, etc.) assume a fixed, known objective, this broadening of the AI foundation to allow for uncertainty over objectives leads to an enormously rich research agenda. Some of the main directions for future work include the following:

- The complexity of modeling real human preferences, including commonalities and differences
  - Learning human preferences from textual sources
  - Semantics of requests/commands/prohibitions/laws
  - Ensuring sufficiently broad priors to avoid model mis-specification
  - Handling plasticity and manipulability of human preferences
  - Inverting real (imperfect, myopic, emotional) human cognition to infer preferences
- Addressing computational limitations in real machines and their effect on provable guarantees
- Theory for multi-human assistance games
  - Mechanism design for avoiding strategic interactions among humans
  - More general and philosophically robust theory of social aggregation

- Theory for (open-source, common-payoff) multi-robot assistance games
- Theory for embeddedness, where the environment includes the agent's computational process
- Development trajectory for scalability and competitive prototype systems

Several other approaches to AI safety were discussed. *Oracles* are AI systems restricted to question-answering, and may be further restricted to systems that output only valid logical or probabilistic inferences. Such systems may still have very high value to humans, yet it may be possible to prove safety properties. The approach of *comprehensive AI services* (Drexler, 2019) extends this idea to an assembly of restricted AI components that collectively have high value when coordinated by humans. Some other approaches including *debate* (Irving et al., 2018) and *iterative amplification* (Christiano et al., 2018) draw on the important insight that verifying/evaluating/comparing specific proposals is strictly easier than devising those proposals in the first place, so that humans can reasonably play the former role while AI systems play the latter role in such a way as to gain human approval. As in assistance games, these approaches assume a run-time flow of preference information from humans to machines. These methods, along with several others, are summarized by Hendrycks et al. (2021) and Hubinger (2020); at least in principle, they could be applied to opaque deep-learning systems.

Russell argued that these considerations of safety would necessitate an approach to developing AI based on semantically well-founded components (with probabilistic programming being a plausible candidate) and a fully verified software stack. He also noted a converse problem: what is to prevent malign or careless actors from deploying powerful but unsafe AI systems? Eventually this would require a systemic redesign of the global digital ecosystem such that only *verifiably safe* systems can execute on standard hardware or be transmitted across networks. Methods such as proof-carrying code (Necula, 1997), with proof-checking implemented in hardware, could make this possible.

Formal Methods for AI Safety (Sanjit Seshia, UC Berkeley)

Seshia discussed formal methods (FM henceforth), mathematical and algorithmic techniques for the modeling, design and analysis of systems, and the challenges and recent progress in using FM for achieving verified AI systems (Seshia et al., 2022). Specification prescribes what the system must do, verification demonstrates that it meets its specification, and synthesis ensures that it meets the specification by design. AI systems today often have data-driven machine learning components, and operate in open-world environments (think autonomous vehicles in traffic) that include humans, creating very high-dimensional input and state spaces that are very difficult to model formally.

Formal verification approaches range from highly automated (simulation-execution based verification) to automated methods requiring some human guidance (model checking) to highly human-intensive (interactive theorem proving for higher-order logic model abstractions).

Seshia contended that AI safety requires that the AI system satisfies its specification, but the central challenge lies in *identifying the right formal specification*, analogous to the alignment problem. Formal specifications correspond to a set of correct behaviors (traces, etc.) or a set of correct implementations. Specifications include models of the agents and their environment in describing a system, and use modeling formalisms including ordinary differential equations (ODEs), finite state machines, and hybrid automata; and express desired properties in formalisms such as propositional logic, first-order logic, temporal logics, and deterministic finite automata (DFAs). Desired formal properties include safety ("nothing bad ever happens"), liveness ("something good eventually happens"), stability, input/output robustness, and integrity.

Seshia discussed three challenges for formal specification of AI systems. The first concerns hard-to-formalize tasks, such as perceptual tasks including object classification, detection, interpreting natural language, etc., for which, even though one may be unable to write a formal specification, one still needs to provide provable guarantees of safety for the system that contains it. The second challenge involves reconciling Boolean versus quantitative specifications, where Boolean specifications are more composable and fit with formal tools, while quantitative specifications are more flexible and fit better with current optimization methods. The third challenge concerns bridging data and traditional formalisms for specification used in FM (as listed above).

In specifying properties of interest, one must choose the right level of abstraction, and consider composability. One can write component-level properties, e.g., the property of a neural network that it maintains monotonicity from input to output. Component properties include robustness, monotonicity, I/O relations, coverage, semantic invariance, and distributional assumptions. An interesting area of development, component level properties of ML systems could include, e.g., counterfactual reasoning. At a system level, properties for ML systems are generally similar to other systems, such as safety, liveness, and stability.

Properties of ML specifications that are not unified or well defined can be formalized by classifying them by purpose. Property classes include robustness (local vs. global, syntactic vs. semantic), input-output relations, monotonicity, fairness, coverage, semantic invariance (e.g. output invariant to geometric transformations), and distributional assumptions & corresponding guarantees (Seshia 2018). Robustness, for example, can be formulated as admissibility, a distance for perturbations, and a target behavior constraint that encompasses the class of robustness in question. This captures the optimization-based formulation of "minimum perturbation, maximum loss induced," rendering the decision and optimization problems analogous, and better characterizing robustness, so that one can predict the circumstances under which perturbations lead to invalid values, and at what scale (Dreossi, et. al. 2019).

Seshia proposed various directions to bridge the gap from logical verification to optimization. One is to approach formalizing data-driven AI systems *at the system level*. One could use temporal logics with quantitative semantics (STL, MTL, Rulebooks, etc.) which would allow for the massaging of verification into optimization; and use specification mining to learn specifications from data (see Vazquez-Chanlatte et al., 2017, 2018; Puranic et al., 2021; Belta et al., 2017, Jha et al., 2017). To model environments, one must inventory what is known and knowable about the environment. Unknowns may be addressed as follows:

- Parameters, with probabilistic programming to define and update distributions
- Behaviors / dynamics, learned from data and direct interaction
- Agents / objects, learned through introspective environment modeling to extract assumptions that the system makes about its world

Seshia described how probabilistic programming languages, e.g., Scenic (Fremont et al., 2019), could be used to model stochastic and dynamic environments of AI systems.

Seshia briefly summarized other research directions for verified AI, including verifying input-output properties of feedforward neural networks; correct-by-construction design of AI systems; oracle-guided learning, learning-based control, neuro-symbolic programming, and runtime assurance. These directions are elaborated on in a recent CACM article (Seshia et al., 2022).

Probabilistic Programming: An Alternate Scaling Route for AI (Vikash Mansinghka, MIT)

Mansinghka described probabilistic programming (PP henceforth), an approach to probabilistic modeling and inference that draws on the full expressive power of Turing-equivalent programming languages (or, in some cases, first-order logic). PPs formalize and automate the implementation of inference, allowing both human users and inference engines to implement custom algorithms for specific models and combine them with generic inference algorithms that, in principle, can handle any model, any computable query, and any data with no further mathematical or algorithm development. Emerging from multiple independent sources in the 1980s and 1990s, probabilistic programming now has its own international conference, with graduate courses at MIT, McGill, Yale, and other top universities internationally. Dozens of probabilistic programming languages (PPLs) have been developed, some of which naturally incorporate deep learning methods, program differentiation (end-to-end gradients), numerical simulators, etc. Several, such as STAN (Carpenter et al., 2017) and Microsoft's infer.net have large user communities and support important applications.

Mansinghka used an MIT-developed PPL, Gen, as an illustrative example (Cusumano-Towner et al., 2019), citing in particular its effective use of compiler technology and user-customizable inference to achieve high efficiency on very challenging applications including computer vision and large-scale database cleaning. In keeping with the overall theme of the workshop, the customizable meta-language for Gen inference programming comes with formal verification of inference soundness (Lew et al., 2020). Gen also automates state-of-the-art estimators for the accuracy of approximate inference algorithms (Cusumano-Towner et al, 2017) and other information-theoretic measures such as entropy and mutual information (Saad et al, 2022), needed for human users (and future automated reasoning engines) to generate accurate, scalable, efficient inference algorithms.

The working hypothesis proposed is that *PP technology, suitably extended, can form an adequate basis for progress in AI towards highly capable or human-level systems* without sacrificing semantic transparency and rigorous theory, while drawing on centuries of results in logic, probability, decision theory, and computation.

The most obvious advantage of PPLs is the universal expressive power of programs, particularly relative to circuit languages such as Bayes nets and neural nets. This allows PPLs to concisely capture complex regularities in environments with large numbers of objects and events, leading to sample-efficient, robust learning. For example, a generative PPL-based computer vision system substantially outperforms the state-of-the-art DenseFusion deep learning system on the YCB benchmark in terms of both accuracy and sample efficiency (Gothoskar et al., 2021). PPLs have been applied to a wide range of important scientific tasks, including climate models, high-energy physics, ecosystem modeling, and Covid-19 epidemiology. Russell's group developed the global seismic monitoring algorithm for the Comprehensive Nuclear-Test-Ban Treaty using the BLOG PPL for the initial model (Arora et al., 2013). Mansinghka also demonstrated that PPL-based perception can naturally handle complex, common-sense environments including household interiors and street scenes with multiple agents.

Contrary to common supposition, PPL models need not be hand-engineered; using generative meta-programs (Saad et al., 2019), it is possible to rapidly explore many possible model structures, handling structural uncertainty with an ensemble of models, yielding both accurate prediction and much more realistic uncertainty quantification on a benchmark time series prediction task. It is also possible to prove the soundness of Bayesian learning procedures for PPLs, an important prerequisite for some formulations of provably beneficial learning.

A number of challenging open problems must be solved for PPLs to yield general-purpose AI capabilities. Although considerable progress has been made in efficient inference via custom algorithms (Cusumano-Towner et al., 2019), generic inference efficiency is paramount: a general-purpose agent cannot continually be reprogrammed for each inference task. One promising approach, that has already achieved state-of-the-art results, is compilation of compositional data-driven MCMC algorithms by static analysis of PPL source code (Lew et al., 2021). This approach has outperformed machine learning baselines and generic PPL MCMC and SMC for inference in probabilistic expert systems that clean and deduplicate databases with millions of records. Deep-learning-based adaptive sampling proposals are another promising direction for importance sampling inference (Le et al., 2017), but the same success has yet to be realized for MCMC and related algorithms, which are fundamentally more effective than importance sampling. To form a basis for PSBAI agents, PPLs must be extended to handle actions, rewards, and decision making. First steps have been taken in this direction (Srivastava et al., 2014; Evans et al., 2021) but much remains to be done.

AI Policy, Standards and Regulation (Gillian Hadfield, U. Toronto)

Hadfield gave a review of the current regulatory landscape for AI, and challenges for the future. Assuming methods to develop provably safe and beneficial AI, she contended that we will need normative (legal/regulatory) infrastructure in place to assure that only safe AI modules are deployed; and a secure, global digital ecosystem such that unsafe AI systems cannot run.

Soft law. Her review of the regulatory landscape began with "soft law," guidelines and principles, as well as industry standards, which some argue will remain the dominant form of AI guidance. Hundreds of sets of AI principles have been adopted since the Asilomar AI Principles were published in 2017 (FLI 2017). But principles and guidelines have no explicit, third-party enforcement. This includes the recent White House blueprint for an AI bill of rights, requiring "independent evaluation and reporting" that systems are indeed safe and effective. Hadfield noted that much of the bill is aligned closely with the PSBAI workshop's goals; and that although not enforceable law, its adoption supports the growing trend of audits and certification provided by third-parties being integrated into the larger ecosystem, illustrated by the growth of start-ups building technical tools to validate against emerging criteria for AI.

Hard law. She then reviewed the rise of hard law itself in contrast to guidance. A new NYC law requires audits of AI and algorithmic systems that drive employment decisions. Several states are enacting mandatory risk management requirements. The EU AI Act, slated for adoption in 2023, which prohibits subliminal manipulation, exploitation of vulnerable groups, social scoring, and some biometrics in law enforcement, is expected to influence hard and soft law in US states. Existing EU laws already require risk management in high-risk fields such as in education, health care, employment, and democratic and judicial process areas. Some workflows and more detailed applications will be specific to industries, in the absence of a unifying framework.

Regulatory ecosystem. Hadfield then discussed the prospect of a regulatory ecosystem, enabling a competitive market for licensed regulatory technologies and services such as compliance tools and audits that verify the validity, implementation and use of AI technologies. This would require governments to define quality and outcome metrics, and require AI developers and users to use licensed regulatory products and services.

Agents as legal entities. To assure AI is cooperative with humans, Hadfield discussed the concept of "AI Passports," noting that amongst humans, "exile" or exclusion from citizenship and membership rights has been an effective penalty in societies throughout history. Continuing the analogy, a centralized authority is empowered to exclude and deport, while there is a corresponding decentralized obligation not to hire or provide benefits for exiled persons. Hadfield's analogous legal framework for AI agents to manage their access to transactions/resources would also require a balance between centralized authority and decentralized enforcement in the context of AI regulation. In another analogy from the law, Hadfield suggested AI agents may enjoy certain personal rights, e.g. to own property, sue and be sued, rights similar to those held by U.S. corporate entities; in this analogy, third-party enforcement plays an important role.

Hadfield pointed to challenges in ensuring an AI agent cannot achieve its objectives without maintaining its license or passport to act, and of forcing an AI agent not to participate without verifying its passport, and concluded with the caution that law itself does not have formal technical definitions; that due process is what fills in the details ex post and decides whether or not the intentionally open-ended 'specification' was violated.

Breakouts

The three breakout sessions were as follows:

1. Well-founded AGI: What are the prospects for achieving AGI using semantically well-defined, compositional, transparent AI systems? What's missing?

2. Safe AI systems: What sorts of safety properties can we prove as AI systems become more capable and complex? What would it take to secure the global digital ecosystem such that unsafe AI systems cannot run?
3. Potential prototype projects: What kinds of prototype systems could be created over the next few years? What tools and environments need to be created and shared?

Each session involved three breakout groups with 15-20 participants each, lasting for two hours per session including 30-45 minutes for plenary report-backs and discussion. Each breakout group had a discussion leader and rapporteur. The following sections summarize the main points from each session.

Breakout 1: Well-founded AGI

What are the prospects for achieving AGI using semantically well-defined, compositional, transparent AI systems? What's missing?

The groups discussed, for each quality (semantically well-defined, compositional, and transparent), its definition, the challenges it poses, and the ways it may help on the path to achieving AGI.

Semantically well-defined systems "make sense" to humans, generally using symbols grounded (tied) to physical concepts. The main challenge with today's AI systems is in defining the semantics of neural network architectures. The choice of grounding is central to safety, because an incomplete or incorrect model of a system or its environment allows an agent, formally validated to operate safely within the model, to be unsafe in the real world. In general, a semantically well-defined system requires a succinct specification of the allowed, desired, and forbidden behaviors of the system according to the model of the system operating within the model of its intended environment. It gives a basis for reasoning about its behavior. In a neural network, we would seek a level of meaning - high-level concepts - beyond a mere input-output distribution; perhaps a separate network could define these high-level concepts.

In discussing symbolic vs quantitative AI, Cyc was cited at the extreme symbolic end of the spectrum; it was noted that it doesn't handle uncertainty, and is not coupled to a learning system, although it could be (cf NELL) .

Compositional systems' modules or agents can be combined like building blocks into larger, more complex systems, ideally passing along to the larger systems the proven properties of their constituent parts. There was discussion of what would be considered "atomic" units or agents, perhaps defined relative to the tasks and environments for which they are built. There was also discussion of examples of deployed AI systems today that are *not* compositional in some form or another: language translation and transcription were given as examples of systems that compose poorly, although that is improving.

In this context the question was posed whether provably safe components might be combined (composed) such that the full system is itself unsafe. The question came up whether systems that are compositional in design are necessarily compositional in implementation.

Transparent systems enable examination of their parts, structure and processes, and most importantly — assuming legibility of the components — allow effective audits to avoid errors and post-mortem analysis when errors do occur. Transparency can give insight into qualities such as compositionality and well-defined semantics. One group discussed the difference between causal explanations, which are valued socially, versus engineering transparency, which allows examining code and data but may explain little in case of audits or error post-mortems.

There were many questions regarding formally specifying AGI. A formal specification for an AI system depends on the level of abstraction. Can we map between the semantics of low- and high-level abstractions? Can we specify high-level abstractions that map to low-level abstractions in a well-defined, meaningful way? Is being semantically well-founded a prerequisite for achieving AGI? Could we ever hope to write down a formal definition of what general intelligence is? Can we approximate it? Should our specifications or definitions evolve over time? It was observed that proving that technology actually is AGI may not be necessary: one could develop technology that suffices for general intelligence, and then prove it's safe (or not), without proving that it is in fact AGI.

Regarding the utility or hindrance provided by these qualities on a path to AGI, some felt that our understanding of compositionality in current AI systems falls short; e.g., Airbus won't use deep learning technology because its compositionality and transparency are not clear to them. However, composable modules would allow swapping out components with newer ones, which would perhaps help on the path to AGI. And pre-AGI systems that are compositional may lead to AGI that is also compositional. But it may be necessary to compromise on these qualities in the pursuit of AGI. It was speculated that AI will quickly exceed our ability to maintain safety, but perhaps in addressing failures we may solve safety issues, as was the case for seatbelts in cars. Nuclear power was invoked as a parallel.

Additional needs and issues discussed regarding well-founded AGI included the following:

- A crisp definition would be useful of what it means for an AI to explain something to a human, just as doctors explain medical issues to patients who have no medical training.
- AI may require a means to model humans' suboptimality.
- Semantically well-founded AI would allow us to better exert governance and control; this should apply to black-box systems as well, such as Waymo's autonomous driving systems, which used to be more compositional, but are increasingly black-box for performance reasons.
- There is a spectrum from black-box testing to proving things about components. For example, many drugs are tested (at a very high cost in resources) to the satisfaction of society in specific applications, although their mechanisms may not be fully understood. Perhaps AI is closer to nuclear science than to pharma.
- How confidence is established depends on the type of system; e.g., Tesla crashes can only look at statistics, while aircraft crashes can identify specific module and protocol failures; this leads to a situation where Tesla crash analysis does not necessarily lead to avoiding repeat failures, whereas aircraft crashes more often do.

The general consensus was that despite the many challenges in developing well-founded AI — semantically well-defined, compositional, and transparent — AGI that is *not* well-founded is highly likely to be unsafe; and that these well-founded qualities may in fact help us on the path to achieving AGI.

However, it is very likely that the necessary technologies for non-well-founded AGI will be developed (and potentially deployed) well before it is possible to build well-founded AGI, presenting a difficult dilemma: it is necessary to both work toward well-founded AGI as quickly as possible, while also working to make non-well-founded approaches as safe as possible to mitigate the potentially serious impacts that are likely to emerge from their deployment.

Breakout 2: Safe AI systems

What sorts of safety properties can we prove as AI systems become more capable and complex? What would it take to secure the global digital ecosystem such that unsafe AI systems cannot run?

Overall the groups concurred that certain properties will be provable as AI systems become more complex, with policy makers likely satisfied with statistical guarantees (a good justification for SMC), and technologists more concerned with provable boolean properties — likely static properties for the time being, but evolving properties important in the future. It was noted that one can still obtain formal guarantees for some parts of a system while other parts are not completely modeled. Computational cognitive science may be called upon for effective models of human behavior. It would help to have structured processes for discovering the properties that we care about.

Examples of hard constraint properties were explored: in synthetic biology, machines are hard-coded not to synthesize gene sequences that are similar to diseases. In networking, the IP protocol disallows sending a packet over the network that does not conform to protocol. These types of constraints require coordination in multi-agent systems: without coordination or resource constraints, all agents tend to overuse the available resources, degrading overall performance.

Much discussion acknowledged the difference between proving properties of a system as defined formally, and measuring observed phenomena of the system once deployed in the real world with humans up to societal scale. Once a system passing a proof-based safety check is released into the (social) wild, it must be tested and monitored as the system's model may be incorrect. Many discussions revolved around how formal methods are currently being used in other fields to evolve specifications over time based on bugs or unintended side effects that are discovered after deployment.

Other properties can lead to untenable results. For example, min-max proofs currently demonstrate that autonomous cars should not even leave the garage. One limitation of properties' expression in control theory is that disturbances must be bounded, but setting the bounds is a challenge. Related to this is the connection between these bounds and a measure of safety; e.g., proving a set of such bounds such that one may conclude that "nothing bad will happen."

It was observed that social media companies such as Meta (formerly Facebook) are already actively monitoring large-scale recommender systems and are faced with securing and maintaining an evolving digital ecosystem. These companies have the challenge of identifying key metrics to monitor the health of the recommender algorithm ecosystem and are faced with translating social science issues into technical terms, e.g., the strength of a link between online speech and offline harms such as linking violent content on social media and violent crime.

Finally, discussion followed regarding the secure digital ecosystem necessary to assure that unsafe systems cannot run. Certification, central to this, is problematic where continuous learning systems are concerned: the means and frequency of re-certification would depend on some assessment, in the context of the application, on risks incurred by the continuous update of parameters.

Breakout 3: Potential prototype projects

What kinds of prototype systems could be created over the next few years? What tools and environments need to be created and shared?

One discussion focused on environments (mainly simulated) in which it would be natural for AI systems to learn to assist humans—the idea being to gradually scale up PBAI towards complex behaviors aiding real human preferences. Some early work on assistance games and human-machine cooperation has used the Overcooked environment in which "chefs" in a highly simplified kitchen prepare meals via preparation of ingredients, cooking, serving, and cleaning up (see, e.g., Carroll at al., 2019). This is roughly analogous to the Taxi world used in many studies exploring new methods for reinforcement learning, but of course it was not designed for the purpose of exploring assistance game algorithms. It would therefore be a good idea to establish one or more benchmark environments to facilitate research on and evaluation of algorithms.

At an intermediate level of complexity there are environments such as Minecraft with much larger state and action spaces and long horizons (on the order of thousands of actions). The BASALT Minecraft challenge (Shah et al., 2021) provides a series of benchmarks of increasing difficulty, aimed eventually at enabling an AI system to help one or more humans in whatever construction activity the humans have decided to engage in. At the moment the temporally extended behaviors required for successful Minecraft construction are beyond the capabilities of typical deep RL systems, although they may be within scope for classical hierarchical planning methods. Recent work (Fan et al., 2022) shows successful reward learning and instruction-following from language-annotated videos of human activities in MineCraft, which is a step towards practical assistance.

Another purpose-built world for learning to assist humans is a game world under development by encultured.ai. The game is roughly comparable to Minecraft but with easier exploration of social and economic elements so as to tap into a wider range of human preferences beyond the purely architectural. It is a game specifically intended to encourage participation by large numbers of humans in order to create a rich learning environment for exploring human preferences and interaction dynamics.

Participants pointed out that there may be a conflict between making environments engaging and making them suitable for embedding useful AI systems. There is also a question of how well human behavior in simulations and games, and the human preferences induced therefrom, carry over to the real world. This is particularly an issue for games such as Overcooked in which there are specific goals to be achieved: in the real world, goals are *never* absolute, but can be overridden depending on circumstances. It also raises the question of how much realism is needed and in which dimensions: while the research community has accumulated plenty of experience with sim2real for robotics, we have essentially none with sim2real for human interaction or preference learning. For example, we have learned that sim2real for robotics requires simulated visual input that engenders the same types and frequencies of errors in visual processing that are engendered by real environments; what sorts of low-level details are necessary in sim2real for preference learning, and what can be abstracted away altogether?

Instead of using simulation or game environments, it may be helpful to acquire and make widely available large datasets of real-world human interactions (i.e., interactions such as travel planning and work activities where the human's real day-to-day preferences are in play). Initially these can be acquired through passive observation of human interactions with ordinary computer systems. As we become more proficient at building active assistance-game solvers—for example in personal digital assistants—these data sets can be acquired in a more directed fashion. Some participants believe that the domain of personal digital assistants would be an excellent challenge task for well-founded AI systems that can reason explicitly about many objects (people, possessions, organizations, locations, etc.) and activities from noisy, partial data and indirect observation.

Safety-related environments can also be adversarial. For example, red-teamers could try to create an AI agent that will deceive humans, while blue-teamers could attempt to detect deception either manually or automatically, while regulators could impose automatically checkable constraints that would make deception more difficult or impossible. Such environments would be an excellent source of demonstrable examples of unsafe AI behaviors and potential mitigations.

Some discussion was devoted to tool development for well-founded AI and formal methods. At present, besides closed-universe PPLs such as infer.net and STAN, which are far too restricted in expressive power to support most AI tasks, the vast majority of PPLs are research prototypes with no widespread user population. Gen is one notable exception: multiple academic courses in the US, Canada, Japan, and Germany have adopted Gen for teaching AI and probabilistic programming. However, the lack of funding for open-source PPL infrastructure has placed PPLs at a significant disadvantage compared to deep learning tools such as TensorFlow, which are supported by large engineering teams, huge data sets, cloud computing services, and special-purpose hardware delivering as much as 10 orders of magnitude more computing power than traditional platforms. Developing and disseminating an industrial-strength, extensible, end-user customizable PPL platform, supporting integration with deep learning models, would be enormously valuable for the broad AI research community and for well-founded AI development in particular. It would also be extremely beneficial to have a broadly available graduate curriculum to accompany the tools, covering both theory and practical application.

In the area of formal methods, there are many tools available—the most powerful being used for tasks such as operating system verification and computer security. For example, the iOS operating system for iPhones and iPads has a fully verified kernel. There are also industrial-strength tools used for control systems in Airbus planes and metropolitan-scale transportation systems. These are somewhat disconnected from the AI community, however, and may not be particularly well adapted to the task of verifying AI systems based on complex perception, learning, and reasoning subsystems. This is clearly an area where integrative research could pay significant dividends.

Lightning talks

Session A

Quantifying misalignment between agents

Shiri Dori-Hacohen (University of Connecticut)

Dori-Hacohen discussed how misalignment can lead to negative, near-term societal impacts. Citing a lack of systematic definition or measures of misalignment, she pointed to two examples: disinformation bots that are aligned with their creator but not the victims, and shopping applications with recommender systems. Drawing on a model of contention that defines controversial by asking “to whom?” and contention-based misalignment by asking “misaligned to whom?” Dori-Hacohen advocated for *population-based misalignment*, defined as the likelihood that two randomly sampled agents from the population hold conflicting goals. She claimed that this explains the phenomenon: social media bots are aligned to only part of the population; and Amazon’s recommender systems’ goal to make money is (potentially) misaligned with customers but not Amazon stakeholders.

Game theory as a framework for thinking about compositionality

Vincent Conitzer (Carnegie-Mellon University; Oxford University)

Conitzer discussed the relationship between multiagent/game-theoretic approaches and compositionality, using the example of GPT3 interacting with DALL-E2, where GPT3 yields an image caption from a prompt that is fed to DALL-E2 to generate an image. The questions that arise from this multi-agent setting include where in the pipeline AI safety problems might appear, and whether current safety measures in GPT3 and DALL-E2 catch problematic prompts. Conitzer posited that one can draw an extensive-form game with imperfect information where DALL-E2 receives a noisy signal about whether GPT3 is safe.

There is already some convergence between the multiagent model and more traditional approaches for composable software systems, for example through formalisms such as program equilibrium where agents are transparent and can read each other's source code.

Implicit bias, counterfactual training, and aligning deep learning systems

Roger Grosse (University of Toronto)

In the context of modern AI systems’ (read: large language models’) capabilities and behaviors such as deception and multi-step reasoning, Grosse advocated for understanding the patterns of generalization of these models to be able to identify which training examples contribute to a given behavior. He mentioned relevant work on decomposing the error in influence functions, and then described Predicting Counterfactual Training (“PCT”) that predicts how the optimal solution changes as one changes the weighting of training examples. Grosse concluded with a discussion of recent insights into what it means to approximate PCT, as well as possible approaches to scaling it up to large language models.

Session B

Simulating humans with Large Language Models

Adam Tauman Kalai (Microsoft Research New England)

Kalai described recent work on a method for using a large language model, such as GPT-3, to simulate responses of different humans in the context of experiments. The work attempted to reproduce well-established economic, psycholinguistic, and social experiments, whose results varied in predictive accuracy of the same experiments run in human populations. GPT-3's completions seemed to predict likely behaviors of participants in the Milgram experiment, where participants were told to administer increasingly powerful electric shocks to another person, who, unbeknownst to the participants, were actually actors pretending to be shocked. Kalai argued that language models are nearly powerful enough for human simulation useful for formulating hypotheses to be tested in risky experiments.

Loopholes and hyper-rationality: AI risk from mechanism-level interactions

Michael Wellman (University of Michigan)

Wellman discussed the risk posed by "hyper-rational" agents from the perspective of economic mechanism design; i.e., agents intelligent and capable beyond the anticipation of mechanism designers. In this setting, the agents may discover "loopholes"—ways to pursue objectives that were not accounted for by the mechanism designers. Wellman illustrated this phenomenon using the example of algorithmic traders manipulating electronically mediated markets.

How OpenAI is planning to do scalable oversight

Jan Leike (OpenAI; Future of Humanity Institute)

Leike described OpenAI's current plans for scalable oversight and the means to measure its efficacy. The plan is based on the insight that, as AI systems perform tasks beyond the ability for humans to evaluate them, AI assistance will enable humans to stay ahead and continue to assure safety and control. It consists of two types of test for LLMs: critiques, where humans rate completions; and dialogue, where humans give text-based feedback to the LLMs. One challenge is that humans tend to be overly trusting of AI systems and miss flaws. OpenAI plans to measure efficacy by (1) manually creating subtly flawed responses using humans; (2) Having other humans label the full pool of responses as flawed or not flawed; and (3) Training the LLM on this data. For increasingly difficult tasks, to use AI assistance to evaluate AI's, one must model the recursive reward of evaluating the AI *assistant*; this leads to a definition of "levels of AI" based on how many levels of assistance are required before humans can reasonably evaluate the assistant.

Towards safer AI

David Krueger (Cambridge University)

In the broad safety issue of goal misgeneralization, Krueger argued that understanding learning curves helps to understand and study generalization; e.g., by studying double descent, grokking, partitioning learning curves into classes of samples, and scaling laws that predict sudden leaps in progress. Krueger discussed the safety-performance tradeoff in several approaches: avoiding instrumental goals, system "myopia" and incentive management, and limiting sensors and actuators. He argued that we must understand how a system may generalize beyond the scope of its training by working around or “hacking” its reward function.

Session C

Evaluating and improving robustness to natural distribution shifts

Aditi Raghunathan (Carnegie Mellon University)

ML systems can fail catastrophically when the test and training distributions differ in some systematic way. Raghunathan described a mathematical characterization of such a distribution shift that enables us to devise robust training algorithms to promote robustness to that specific class of shifts. However, the resulting robust models show limited gains on shifts that do not admit the structure they were specifically trained against. Naturally occurring shifts are hard to predict a priori and intractable to mathematically characterize when occurring in the wild. Raghunathan discussed how to estimate the performance of models under natural distribution shifts from small to catastrophic. Obtaining ground truth labels is expensive and requires a priori knowledge of time and type. With so-called “agreement-on-the-line,” they effectively predict performance under distribution shift from unlabeled data alone. A promising avenue for improving robustness to natural shifts leverages representations pre-trained on diverse data. Via theory and experiments, Raghunathan found that the de facto fine-tuning of pre-trained representations does not maximally preserve robustness. She described two simple alternate fine-tuning approaches that substantially boost robustness to natural shifts.

Building certifiably safe and correct large-scale autonomy

Chuchu Fan (MIT)

Fan argued that learning-based methods in building autonomous systems can be extremely brittle in practice and are largely not designed to be verifiable. She described several recent efforts that combine ML with formal methods and control theory to enable the design of provably dependable and safe autonomous systems, as well as techniques to generate safety certificates and certified control for complex autonomous systems, even when the systems have a large number of agents (e.g., thousands of quadcopters) and follow nonlinear and nonholonomic dynamics (e.g., navigating aerial traffic in a city).

Safe Learning - A Perspective from Control

Claire Tomlin (UC Berkeley)

Tomlin described safety in control (directing the behavior of dynamic, engineered systems) today as largely focused on using reachability analysis wherein challenges come primarily from computing these reachable sets, limited to one or just a few agents. Safe learning is implemented by measuring disturbances in real-time to recompute these reachable sets. In automating more complex, hierarchical platforms like autonomous cars, aircraft or air-traffic control, the original guarantees and training data used to train for safe control of subsystems do not naturally remain valid at the full system-of-systems level. Provable safety in this context requires us to enable predictable, safe, and high-confidence interactions between humans and the machines that work with them - even where the specification is unknown, where the distribution of future data will not necessarily follow the distribution of the past, operating in larger scale settings, with teams of humans and robots. One guiding principle in this work will be to focus on safety of the holistic systems, not just their components (Tomlin, 2021).

Deep Reinforcement Learning with Formally Verified Safety

Swarat Chaudhuri (UT Austin)

Chaudhuri observed that recent approaches to the formal verification of deep learning systems decouple learning and verification: one first trains a neural network with state-of-the-art deep learning techniques, then verifies it using formal methods. Unfortunately, such decoupling poses a fundamental issue: a model whose training objective does not include a correctness property may or may not satisfy the property after training. Over the last two years, Chaudhuri's group has built a body of deep learning and neuro-symbolic learning techniques that respond to this challenge. They have developed several deep reinforcement learning algorithms that invoke formal verification from inside the learning loop and guarantee provable safety either at convergence or during training.

Symbols as a Lingua Franca for Explainable & Advisable AI Systems

Subbarao (Rao) Kambhampati (Arizona State University)

Despite recent LLM's power to learn their own representations, Kambhampati argued that their inscrutability leads to problems in their safety and ability to interact with humans. Neuro-symbolic approaches are often motivated by (i) symbols as a *lingua franca* for human-AI interaction and (ii) symbols as system-produced abstractions used by an AI system in its internal reasoning. It is unclear whether AI systems will need to use symbols in their internal reasoning to achieve general intelligence capabilities, but either way, they will need symbols for human-AI interaction. In many human-designed domains, humans use explicit (symbolic) knowledge and advice -- and expect machine explanations in kind. Kambhampati advocated for research directions that enable symbolic human-AI interaction.

Truthfulness, interpretability, and emergence

Jacob Steinhardt (Berkeley)

Steinhardt described recent work in the area of "mechanistic interpretability", which aims to understand the latent structure in models' hidden states. Using this perspective, he demonstrated ways to make language models produce more truthful outputs, as well as better understand emergent model behaviors, such as the "grokking" phenomenon first observed by Power et al. (2021).

Session D

On the role of mechanism design in recommender ecosystems

Craig Boutilier (Google Research; U of Toronto)

Recommender systems (RSs) lie at the center of complex ecosystems, involving large numbers of users, content providers or vendors, advertisers and even competing platforms, whose behaviors are driven by their incentives or preferences for RS-induced outcomes. The resulting interactions can generate complex dynamics which, in turn, impacts the ability of the RS to act in the best interests of any particular actor or implement tradeoffs w.r.t. the interests of different actors. The design of RSs in such settings has received relatively scant attention. Boutilier briefly illustrated examples of such interactions (see, e.g., Mladenov et al., 2020) and discussed the use of mechanism design (MD) —and adjacent areas, such as preference elicitation, behavioral economics, reinforcement learning, etc.—as a means to ensure RSs have positive societal impacts. He described a number of research challenges that must be addressed to bring MD to bear on recommender ecosystems.

Human-Aligned Reinforcement Learning: A multiobjective approach

Richard Dazeley (Deakin University - Australia)

Dazeley discussed Human-Aligned Reinforcement Learning (HARL) in the context of the open-ended and black-box nature of autonomous AI systems, which, integrated into human environments, continue to raise concerns from governments, industry, researchers, and civil society. To assure that the behavior of these systems remains beneficial to humanity, Dazeley described HARL, which investigates this nexus between autonomous RL-based systems and humans through the development of approaches that align them in mixed domains. HARL focuses on how an agent can learn how to safely and ethically interact with people, as well as explain how these behavioral constraints affect their behavior. They believe a socially integrated autonomous agent can best combine these components through a multiobjective approach.

Aligning Recommender Systems

Jonathan Stray (Berkeley CHAI)

Recommender systems are among the largest deployed AI systems. Stray argued that if we can't align RSs, we probably can't align more advanced societal-scale systems. He gave examples demonstrating that user preferences cannot be determined from behavior alone, and argued for an assistance game paradigm, where the recommender asks the user for more information. The simplest way to do this is to use evaluative survey measures, such as “was this item valuable to you?” or a wide variety of more general well-being measures. While not all users can be surveyed, it is possible to use the answers from a subset of users to predict to some degree what other users would have answered, and to use these predictions as ranking signals. This technique is already widely used in industry, however surveys are typically only performed at a single point in time and not necessarily on the measures that AI safety researchers might worry about. The next step is to use longitudinal surveys, which results in a very long horizon, multi-user reinforcement learning problem. Stray described a large-scale alignment experiment currently underway in collaboration with Facebook to develop and test this methodology, attempting to modify the Facebook News Feed to optimize for a well-being survey measure over multiple months. He argued for the need for independent funding sources for such experiments.

Preventing undesirable behavior of intelligent machines

Emma Brunskill (Stanford)

Brunskill spoke on learning from limited samples to make good decisions robustly. Limiting the scope to known undesirable behavior, she argued this is essentially constrained optimization, for which a Seldonian approach applies, given historical data to assure a solution satisfies some constraints over that data, i.e., a batch RL approach with safety constraints. She gave the example of insulin diabetes management where side effects are a major concern. In this case, one collects data under some default known drug in order to analyze the performance of a new drug and the likelihood that the new drug will satisfy the side effect constraints. This safe approach performs as well as a pure performance-optimizing approach. Brunskill then extends the concept to an infinite policy space, modifying the algorithm so that it avoids merely resulting in "no solution found." There is a large body of related work on using offline RL to increase the robustness of performance, similar to robust MDPs.

Appendix B: Workshop Attendees

The workshop had 54 attendees, listed here with site links, area specialty, and general workshop area.

First Name	Last Name	Personal Website URL	Speciality	Workshop Area
Andrea	Bajcsy	https://people.eecs.berkeley.edu/~abajcsy/	safety meets learned models of humans	General AI Safety
Craig	Boutilier	https://research.google/people/CraigBoutilier/	aligning rec systems	General AI Safety
Sam	Bowman	https://cims.nyu.edu/~sbowman/	Scalable oversight; large language models	General AI Safety
Emma	Brunskill	https://cs.stanford.edu/people/ebrun/index.html	Computer Science, RL, Theory, Education	General AI Safety
Ilaria	Canavotto	https://sites.google.com/view/ilariacanavotto/		General AI Safety
Swarat	Chaudhuri	https://www.cs.utexas.edu/~swarat/		Formal Methods
Vince	Conitzer	https://users.cs.duke.edu/~conitzer/	Computer Science, Econ, Philosophy, PPE	General AI Safety
Anthony	Corso	https://anthonylcorso.com/		General AI Safety
David	Danks	https://www.daviddanks.org/		General AI Safety
Richard	Dazeley	https://www.deakin.edu.au/about-deakin/people/richard-dazeley	Human-aligned Reinforcement Learning (Safe, Ethical and Explainable RL), Multi-objective RL	General AI Safety
Shiri	Dori-Hacohen	https://shiri.dori-hacohen.com/		Formal Methods
Anca	Dragan	http://people.eecs.berkeley.edu/~anca/	Computer Science	General AI Safety
David	Duvenaud	http://www.cs.toronto.edu/~duvenaud/	AI	General AI Safety
Chuchu	Fan	https://chuchu.mit.edu/		Formal Methods
Jaime	Fisac	https://ece.princeton.edu/people/jaime-fernandez-fisac	Safe Robotics	General AI Safety

Marion	Fourcade	https://sociology.berkeley.edu/faculty/marion-fourcade	Sociology: Morality in the Digital Economy	AI Governance/Econ/Policy
Iason	Gabriel	https://www.linkedin.com/in/iason-gabriel/		General AI Safety
Jemin	George	https://www.linkedin.com/in/jemin-george/		
Mohammad	Ghavamzadeh	https://mohammadghavamzadeh.github.io/		Symbolic AI/PPLs/etc.
Roger	Grosse	https://www.cs.toronto.edu/~rgrosse/	AI	General AI Safety
Gillian	Hadfield	https://www.law.utoronto.ca/faculty-staff/full-time-faculty/gillian-hadfield	Law, Econ, Tech & Society	AI Governance/Econ/Policy
Dylan	Hadfield-Menell	http://people.csail.mit.edu/dhm/	CIRL	General AI Safety
Joe	Halpern	https://www.cs.cornell.edu/home/halpern/		Formal Methods
Nick	Hay	https://www.linkedin.com/in/nick-hay-801855182/		General AI Safety
Wes	Holliday	https://philosophy.berkeley.edu/people/detail/348	Logic & Social Choice Theory	General AI Safety
John	Horty	http://www.horty.umiacs.io/		
Susmit	Jha	https://susmitjha.github.io/		Formal Methods
Sizhe (Jessie)	Jin	US Army, Safety Engineering	System Safety	Policy and Regulations
Adam	Kalai	https://www.microsoft.com/en-us/research/people/adum/	ML theory and code generation	Formal Methods
Rao	Kambhampati	https://rakaposhi.eas.asu.edu/		Formal Methods
David	Krueger	https://www.davidscottkrueger.com/	Ai safety	General AI Safety
Joel	Leibo	http://www.jzleibo.com/		General AI Safety
Jan	Leike	https://jan.leike.name/	value alignment lead	General AI Safety
Tengyu	Ma	http://ai.stanford.edu/~tengyuma/		Formal Methods
Mark	Nitzberg	https://thedecisionlab.com/author/mnitzberg	Computer Science	General AI Safety
George	Pappas	https://www.georgejpappas.org/	verification of ML, among other things	Formal Methods
Ariel	Proccacia	http://procaccia.info/	game-theoretic aspects of prefernce learning	General AI Safety

Aditi	Raghunathan	https://www.cs.cmu.edu/~aditirag/	robustness in SL and RL	Formal Methods
Nicholas	Renninger			Formal Methods
Stuart	Russell	https://people.eecs.berkeley.edu/~russell/	Computer Science	General AI Safety
Dorsa	Sadigh	https://dorsa.fyi/	Computer Science	General AI Safety
Bart	Selman	https://www.cs.cornell.edu/selman/		Formal Methods
Sanjit	Seshia	https://people.eecs.berkeley.edu/~sseshia/	Computer Science, Formal Methods, Verification	Formal Methods
Ameesh	Shah	https://ameesh-shah.github.io/		
Dawn	Song	https://people.eecs.berkeley.edu/~dawnsong/		Formal Methods
Diane	Staheli	https://www.linkedin.com/in/dianestaheli/	DOD Responsible AI	
Jacob	Steinhardt	https://jsteinhardt.stat.berkeley.edu/	ML	General AI Safety
Jonathan	Stray	http://jonathanstray.com/	Computer Science, Recommender Systems, Algorithmic Media	AI Governance/Econ/Policy
Claire	Tomlin	https://people.eecs.berkeley.edu/~tomlin/	safety + learning + control	Formal Methods
Ufuk	Topcu	https://www.ae.utexas.edu/people/faculty/faculty-directory/topcu		Formal Methods
Wilfredo (Wil)	Vega	US Army, Safety Engineering	Safety Engineering	General AI Safety
Adrian	Weller	http://mlg.eng.cam.ac.uk/adrian/	ML, HCAI	General AI Safety
Mike	Wellman	https://cse.engin.umich.edu/personnel/wellman-michael	comp. market mechanisms and game-theoretic reasoning methods	General AI Safety
John	Zysman	https://brie.berkeley.edu/john-zysman	AI Governance	AI Governance/Econ/Policy

Appendix C: Workshop Schedule

The workshop included 4 overview talks, 4 breakout sessions, and 17 lightning talks over 3 days.

	Friday October 7 At Residence Inn Marriott - 3rd Floor Ballroom	
	Activity	**Details**
11:00 AM	Registration (1hr)	Marriott 3rd Floor -- Foyer
11:15 AM		
11:30 AM		
12:00 PM	Lunch	Marriott 3rd Floor -- Foyer
12:15 PM		
12:30 PM		

1:00 PM	Introductions (1hr)	Welcome message (Stuart Russell);
		Self-introductions by all attendees;
		Overview of NSF Convergence Accelerator Program (Jemin George);
		Overview of UMD w/s on Ethical Design of AI (David Danks)

1:15 PM		
1:30 PM		
2:00 PM	Overview Talk: General AI Safety (1hr)	Stuart Russell
2:30 PM		
2:45 PM		
3:00 PM	Coffee/Tea & Snack Break (30min)	
3:30 PM	Overview Talk: Formal Methods (1hr)	Sanjit Seshia
4:00 PM		
4:15 PM		
4:30 PM	Overview Talk: PPLs/symbolic/well-founded AI (1hr)	Vikash K. Mansinghka
5:00 PM		
5:30 PM	Short Break (15min)	
5:45 PM	Lightning Talks: Session A (45min)	Shiri Dori-Hacohen Vincent Conitzer Roger Grosse
6:00 PM		

6:30 PM	Dinner @ Marriott, informal reception in dining area & lobby	Marriott 1st Floor -- Dinning Room Area (Ashby Room) **Dinner served approx. 6:45pm**
7:00 PM		
7:30 PM		
8:00 PM		
8:30 PM		

	Saturday October 8 **At Residence Inn Marriott - 3rd Floor Ballroom**	
	Activity	**Details**
7:30 AM		
8:00 AM	Breakfast @ The Marriott	Marriott 3rd Floor -- Foyer
8:30 AM		
9:00 AM	Lightning Talks: Session B (1 hr)	Adam Kalai Michael Wellman Jan Leike David Krueger
9:30 AM		
9:45 AM		
10:00 AM	Breakout Topic #1 Well-founded AGI?	What are the prospects for achieving AGI using semantically well-defined, compositional, transparent AI systems? What's missing? *Self-scheduled 15 min break in between*
10:30 AM		
11:00 AM		
11:15 AM		
11:30 AM	Reassemble and Discuss (30min)	
12:00 PM	Lunch	Marriott 3rd Floor -- Foyer
12:15 PM		
12:30 PM		
1:00 PM		
1:15 PM	Lightning Talks: Session C (1.5hr)	Aditi Raghunathan Chuchu Fan Claire Tomlin Swarat Chaudhuri Rao Kambhampati Jacob Steinhardt
1:30 PM		
2:00 PM		
2:30 PM		
2:45 PM	Breakout Topic #2 Safe AI systems?	Start with 15 min break What sorts of safety properties can we prove as AI systems become more capable and complex? What would it take to secure the global digital ecosystem such that unsafe AI systems cannot run?
3:00 PM		
3:30 PM		
4:00 PM		
4:15 PM	Reassemble and Discuss (45 min)	

4:30 PM		
5:00 PM	Overview Talk: AI Policy, Regulation, and Standards	Gillian Hadfield
5:30 PM		
5:45 PM		
6:00 PM	Shuttle from Marriott to Faculty Club for Dinner	**Please meet in 1st Floor Lobby no later than 6:10pm for shuttle**

	Sunday October 9 **At Soda Hall (UC Berkeley Campus)**	
	Activity	**Details**
8:00 AM		
8:30 AM	Shuttle from Marriott to Soda Hall	**Please meet in 1st Floor Lobby no later than 8:40am for shuttle**
9:00 AM	Lightning Talks: Session D (1 hr)	Craig Boutilier Richard Dazeley Jonathan Stray Emma Brunskill
9:30 AM		
9:45 AM		
10:00 AM	Breakout Topic #3 Potential prototype projects	What kinds of prototype systems could be created over the next k years? What tools and environments need to be created and shared? *Self-scheduled 15 min break in between*
10:30 AM		
11:00 AM		
11:15 AM	Reassemble and Discuss (45min)	
11:30 AM		
12:00 PM	Closing Remarks (15min)	
12:15 PM	Lunch @ Wozniak Lounge (Soda Hall)	
12:30 PM		
1:00 PM	Shuttle back to Marriott for luggage	**Please meet in 1st Floor Lobby no later than 1:10pm for shuttle**
1:15 PM		
1:30 PM	End of workshop	

References

Arora, N. S., Russell, S. J., and Sudderth, E. (2013). NET-VISA: Network processing vertically integrated seismic analysis. *Bulletin of the Seismological Society of America*, 103, 709–729.

Belta, C., Yordanov, B., and Gol, E. A. (2017). *Formal Methods for Discrete-Time Dynamical Systems*. Vol. 15. Cham: Springer International Publishing.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). STAN: A probabilistic programming language. *Journal of Statistical Software*, 76, 1–32.

Carroll, M., Shah, R., Ho, M. K., Griffiths, T., Seshia, S., Abbeel, P., and Dragan, A. (2019). On the utility of learning about humans for human-AI coordination. In Proc. NeurIPS-19.

Carter, B., Jain, S., Mueller, J., and Gifford, D. (2021). Overinterpretation reveals image classification model pathologies. In *Proc. NeurIPS-21*.

Christiano, P., Shlegeris, B., and Amodei, D. (2018). Supervising strong learners by amplifying weak experts. arXiv:1810.08575.

Cusumano-Towner, M. F., Saad, F. A., Lew, A. K., and Mansinghka, V. K. (2019). Gen: A general-purpose probabilistic programming system with programmable inference. In *Proc. PLDI-19*.

Dreossi, T, Ghosh, S., Sangiovanni-Vincentelli, A., and Seshia, S. (2019). A formalization of robustness for deep neural networks. In *VNN '19*.

Drexler, K. E. (2019). Reframing superintelligence: Comprehensive AI services as general intelligence. Technical Report #2019-1, Future of Humanity Institute, University of Oxford.

Evans, O., Stuhlmüller, A., Salvatier, J., and Filan, D. (2021). *Modeling Agents with Probabilistic Programs*. Agentmodels.org.

Fan, L., Wang, G., Jiang, Y., Mandlekar, A., Yang, Y., Zhu, H., Tang, A., Huang, D.-A., Zhu, Y., Anandkumar, A. (2022). MineDojo: Building open-ended embodied agents with internet-scale knowledge. In *Proc. NeurIPS-22*.

Future of Life Institute (FLI 2017). AI Principles. Developed at the 2017 Beneficial AI Conference held at Asilomar, California.

Gershman , S. J., Horvitz , E. J., and Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, 349(6245), 273-278.

Gleave, A., Dennis, M., Kant, N., Wild, C., Levine, S., and Russell, S. (2020). Adversarial policies: Attacking deep reinforcement learning. In *Proc. ICLR-20*.

Gothoskar, N., Cusumano-Towner, M., Zinberg, B., Ghavamizadeh, M., Pollok, F., Garrett, A., Tenenbaum, J., Gutfreund, D., and Mansinghka, V. (2021). 3DP3: 3D scene perception via probabilistic programming. In *Proc. NeurIPS-21*.

Hadfield-Menell, D., Dragan, D., Abbeel, P., and Russell, S. (2016). Cooperative inverse reinforcement learning. In *Proc. NeurIPS-16.*

Hadfield-Menell, D., Dragan, D., Abbeel, P., and Russell, S. (2017). The off-switch game. In *Proc. IJCAI-17*.

Hendrycks, D., Carlini, N., Schulman, J., and Steinhardt, J. (2021). Unsolved problems in ML safety. arXiv:2109.13916.

Hubinger, E. (2021). An overview of 11 proposals for building safe advanced AI. arXiv:2012.07532.

Irving, G., Christiano, P., and Amodei, D. (2018). AI safety via debate. arXiv:1805.00899.

Jha, S. and Seshia, S. (2017). A theory of formal synthesis via inductive learning. *Acta Informatica* 54, no. 7: 693-726.

Koller, D., McAllester, D. A., and Pfeffer, A. (1997). Effective Bayesian inference for stochastic programs. In *Proc. AAAI-97*.

Kwiatkowska, M., Norman, G., Parker, D., Qu, H. (2010) Assume-Guarantee Verification for Probabilistic Systems. In: Esparza, J., Majumdar, R. (eds) Tools and Algorithms for the Construction and Analysis of Systems. TACAS 2010. Lecture Notes in Computer Science, vol 6015. Springer..

Le, T. A., Baydin, A. G., and Wood, F. (2017). Inference compilation and universal probabilistic programming. In *Proc. AISTATS-17*.

Lew, A. K., Cusumano-Towner, M., Sherman, B., Carbin, M., and Mansinghka, V. K. (2020). Trace types and denotational semantics for sound programmable inference in probabilistic languages. In *Proc. POPL-20*.

Mladenov, M., Creager, E., Ben-Porat, O., Swersky, K., Zemel, R., Boutilier, B. (2020). Optimizing long-term social welfare in recommender systems: A constrained matching approach. In *Proc. ICML-20.*

Necula, G. C. (1997). Proof-carrying code. In *Proc. POPL-97*.

Power, A., Burda, Y., Babuschkin, H., and Misra, V. (2022). Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint* arXiv:2201.02177.

Puranic, A., Deshmukh, J., and Nikolaidis, S. (2021). Learning from demonstrations using signal temporal logic. In *PMLR 2021*.

Russell, S., and Subramanian, D. (1995). Provably bounded-optimal agents. *Journal of Artificial Intelligence Research*, 2.

Russell, S. (2019). *Human Compatible: AI and the Problem of Control* . New York: Viking.

Saad, F. A., Cusumano-Towner, M. F., Schaechtle, U., Rinard, M. C.. and Mansinghka, V. K. (2019). Bayesian synthesis of probabilistic programs for automatic data modeling. In *Proc. POPL-19.*

Saad, F. A., Cusumano-Towner, M. F., and Mansinghka, V. K. (2022). Estimators of entropy and information via inference in probabilistic models. *In Proceedings of the 25th International Conference on Artificial Intelligence and Statistics,* PMLR 151:5604-5621.

Seshia, S., Desai, A., Dreossi, T., Fremont, D., Ghosh, S., Kim, E., Shivakumar, S., Vazquez-Chanlatte, M., and Yue, X. (2018). Formal specification for deep neural networks. In: *ATVA 2018*. LNCS, vol 11138. Springer, Cham.

Seshia, S., Sadigh, D., and Sastry, S. (2022). Towards verified artificial intelligence. In *Communications of the ACM 65:7* pp 46–55.

Shah, R., Wild, C., Wang, S., Alex, N., Houghton, B., Guss, W., Mohanty, S., Kanervisto, A., Milani, S., Topin, N., Abbeel, P., Russell, S., and Dragan, A. (2021). The MineRL BASALT competition on learning from human feedback. arxiv.org/abs/2107.01969.

Srivastava, S., Russell, S., and Ruan, P. (2014). First-order open-universe POMDPs. In *Proc. UAI-14*.

Stray, J., Halevy, A., Assar, P., Hadfield-Menell, D., Boutilier, C., et. al. (2022) Building Human Values into Recommender Systems: An Interdisciplinary Synthesis. http://arxiv.org/abs/2207.10192

Tomlin, C. (2021). Safe Learning in Robotics. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining.*

Vazquez-Chanlatte, M., Jha, S., Tiwari, A., Ho, M., and Seshia, S. (2017, 2018). Learning task specifications from demonstrations. In *Proc. NIPS '18.*

Wang, T. T., Gleave, A., Belrose, N., Tseng, T., Miller, J., Dennis, M., Duan, Y., Pogrebniak, V., Levine, S., and Russell, S. (2022). Adversarial policies beat professional-level Go AIs. *arXiv preprint* arXiv:2211.00241.

Zhuang, S. and Hadfield-Menell, D. (2020). Consequences of misaligned AI. In *Proc. NeurIPS-20.*