JOHNS HOPKINS
INSTITUTE *for*
ASSURED AUTONOMY

IAA Report on AI Safety and Assurance

I. Overview

I-Jeng Wang and Aryeh Englander

Introduction

Artificial intelligence has been advancing rapidly in recent years and is becoming increasingly ubiquitous and impactful. People around the world use AI-based search engines, image recognition, and machine translation on a daily basis. AI is used extensively in medical diagnosis, the financial markets, business analytics, and manufacturing. Many cars have at least a limited form of AI-powered self-driving ability. Closer to the cutting edge of research, there have been impressive and well-publicized advances in natural language generation, game playing, and medical research. (See e.g., [1]–[3].)

However, researchers have discovered that these new capabilities come with many new or exacerbated challenges. For example, tiny changes to an image can trick classification algorithms into misclassifying the image, and subtly mis-specifying the algorithm's goals can lead to surprising and unwanted behavior. Modern AI algorithms are often very opaque to human understanding, they often have difficulty generalizing to environments that they have never encountered before, and they can be extremely difficult to test for compliance with regulations. There are also many challenges related to security, human-machine interactions, ethics, and governance. These problems have stymied the advance of potentially very beneficial AI applications, and they have led to increased worries about the potential damage that AI applications may cause.

The purpose of this report is to provide a summary of the current state of AI safety and assurance research, including descriptions of the different types of challenges as well as current research directions. The report is intended for three primary audiences: Institute for Assured Autonomy leadership and project managers; policy makers with general AI knowledge; and technical and academic communities across the computer science, engineering, and technology policy domains.

Scope and Terminology

This report is primarily focused on autonomous systems that utilize machine learning (ML), since machine learning is the key driver behind many of the most difficult challenges facing AI and autonomous systems more broadly. As a secondary focus, we will discuss autonomous systems and AI/ML systems more broadly as they relate to or are similar to the challenges faced by ML-enabled autonomous systems. (See Figure 1.)

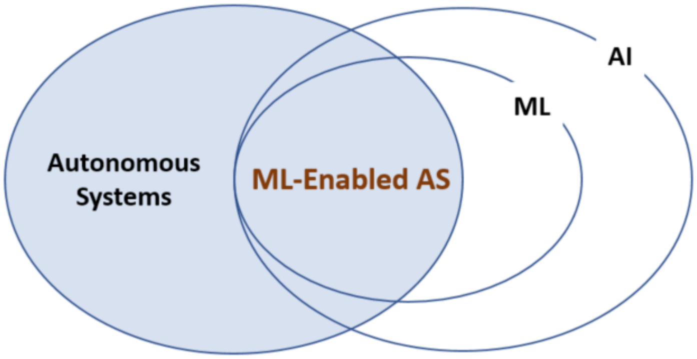

Figure 1: This report focuses specifically on ML-enabled autonomous systems. We do not focus as much on autonomous systems or AI and ML systems more broadly, except as they relate to the ML-enabled autonomous systems challenges we discuss.

In this report we will use the following general definitions:

- Autonomous system (AS): A system that can make decisions independently or with minimal supervision from human operators.[4]
- Artificial intelligence (AI): A software system that can reason in ways comparable to the way that humans think, or comparable to some idealized standard of rational thinking.[1][5]
- Machine learning (ML): A set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty.[6]
- Critical system: A system whose failure may lead to injury or loss of life, damage to the environment, unauthorized disclosure of information, or serious financial losses.[7]
- Safety-critical system: A system whose failure may result in injury, loss of life, or serious environmental damage.[7]
- Assurance: Justified confidence that the system will perform as expected.[4]

Note that even though we mostly focused on ML, we will nevertheless continue to use the term “AI safety” because that is the term that is commonly used.

[1] Unfortunately, the term “artificial intelligence” is somewhat ambiguous and is used by different authors to mean different things. Additionally, what is considered “comparable to the way that humans think” changes as computers become more capable and people stop considering certain capabilities as something human-like. More recently, some people have started using the term “artificial intelligence” to refer almost entirely to machine learning, to the exclusion of other types of “good old fashioned AI” (GOFAI).

Also note that this report does not exclusively focus on technical challenges and solutions, but on all challenges related to critical and especially safety-critical AI-enabled systems. This includes issues of ethics and accountability, governance, and risk analysis, among others.

Background

Artificial Intelligence and Machine Learning[2]

Earlier periods of major AI research focused primarily on symbolic reasoning, which uses high-level “symbols” to represent human knowledge and reasoning in ways that computers can process. The “expert systems” perspective, which uses inferences based on large numbers of special-purpose rules derived from experts, achieved great commercial success in the 1980s. However, expert systems failed to scale up to very complex applications, primarily due to an inability to deal well with uncertainty or to learn from experience. This led to a period of “AI winter” when enthusiasm and funding dried up. Although other types of AI have since taken over the forefront of AI research, these types of symbolic “good old-fashioned AI” (GOFAI) are still widely used in many applications.

Other types of algorithms that are often considered under the label of “AI” include various types of search algorithms, evolutionary algorithms, constraint optimization algorithms, and rule-based planning algorithms, among many others. AI is often also considered to encompass the areas of automated planning, natural language processing, computer vision, and robotics, regardless of the algorithm employed.

More recently, the AI field has been dominated by machine learning algorithms. Traditionally, ML has been divided into three categories: supervised learning, unsupervised learning, and reinforcement learning (RL).

Supervised learning refers to when the AI learns how to map inputs to appropriate outputs by observing labeled input-output pairs. For example, supervised image classification algorithms learn to correctly categorize images by observing thousands or millions of images together with their associated categories. Unsupervised learning refers to finding patterns in data without explicit feedback, for example by observing that certain images have characteristics that cluster together, without being given any labels for that cluster. Finally, reinforcement learning refers to learning a course of action from a series of rewards or punishments, for example learning how to play a game by playing it many times and receiving rewards for winning.

There are many different approaches to machine learning. Some of the major approaches include regression analysis, clustering techniques, support vector machines, and probabilistic and causal inference algorithms.

[2] This section and the next are primarily based on [5].

Parallel to these advances, work on robotics has also accelerated in recent years. The early but very limited semi-autonomous robots of the 1940s and 50s have been replaced with robotic industrial arms, dexterous grasping arms trained with deep learning techniques, self-driving cars, robotic delivery drones, surgical robots, and multi-limbed robotic creatures that can operate in a wide range of environments.

Artificial Neural Networks and Deep Learning

One of the most successful machine learning techniques has been artificial neural networks (ANNs). The idea of ANNs started in the early days of AI research, and was inspired by biological neural networks in the brain. At a basic level, ANNs are networks of interconnected nodes or “neurons,” each of which is assigned a set of numerical weights. When provided with an input (typically a large vector or matrix), each node applies a function that uses the weights to output a value. The value is then passed on to the next node or is used as one of the outputs of the system. During training, the system as a whole is assigned an objective function that is used to “grade” how well the system did at accomplishing its objective, and then that grade is used to tweak all of the weights in the system in order to better achieve the objective on the next attempt. This process is repeated again and again (possibly millions of times) until the system converges on a set of weights that achieves the best output for its assigned objective.

Neural networks turned out to be impractical for anything really useful when they were first proposed, and they languished as a backwater of AI research until the popularization of the back-propagation algorithm in 1986.[3] Deep neural networks, also known as deep learning (DL) algorithms, really took off in 2012, when a DL-based image recognition algorithm far surpassed all previous algorithms in the influential ImageNet competition.[8] DL involves many layers of neural nets connected to each other, with the total number of parameters sometimes numbering in the billions for some of the largest networks. The invention of the internet and the subsequent availability of massive amounts of “big data,” along with greatly increased computing resources using GPUs, has allowed DL algorithms to dramatically increase in capability and applicability over the past decade.

DL has exploded into virtually every area of AI research, with very impressive results in many fields including image recognition, natural language processing, machine translation, and automated planning. DL has also started to be used for reinforcement learning, including for game playing, robotics, and a large number of other applications.

[3] Back propagation had been independently invented several times before that, but it did not become widely known until the 1986 publication of a nonmathematical presentation of the idea in *Nature*.

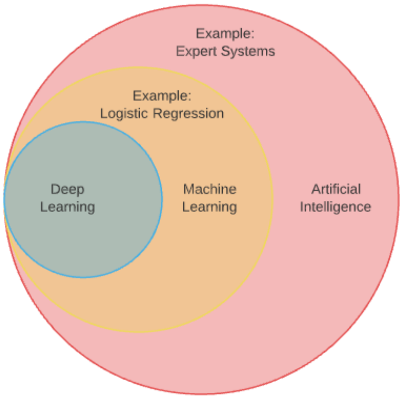

Figure 2: Deep learning is a type of machine learning, which is a subset of artificial intelligence more broadly. Figure adapted from [9].

AI Safety and Assurance Research Communities

As we will discuss in detail in the rest of this report, these advances have led to numerous new or increased challenges in AI safety and assurance. Fortunately, several different research communities have begun to research these new challenges from different angles. The computer science and machine learning communities have approached the issues primarily from a technological framework, with a focus on technical algorithmic challenges. The systems engineering and safety engineering communities have tried to build on their existing techniques to find ways in which regulators can assure that AI algorithms are functioning correctly, and to find ways to adequately conduct testing, evaluation, verification, and validation (TEVV) for AI-enabled autonomous systems. The ethics and policy communities have also been heavily involved in related research from their own perspectives.

In addition to these research communities, in recent years the subject of AI Safety has received a lot of attention from researchers interested in the long-term future of AI and the possibility of global catastrophic risks caused by very advanced AI. Research into these topics started primarily within certain internet-based communities, but has since expanded into a rapidly growing body of interdisciplinary research.[4]

While initially these disparate research communities were only vaguely aware of each others' work, more recently members of all these communities have begun hosting joint conferences and interacting on a regular basis. Organizations such as the Johns Hopkins Institute for Assured Autonomy (IAA), the Assuring Autonomy International Programme (AAIP), and the Consortium on the Landscape of AI Safety (CLAIS) are attempting to systematize and integrate these disparate fields. This report can be viewed as an important part of this systematization and integration.

[4] For a fascinating history of this research community, see [10].

Key Challenges

Increased Capabilities, Increased Challenges

In the past, autonomous and AI-enabled applications were constrained to operating within carefully controlled settings. Before using such a system in the real world, designers could sit down with domain experts and brainstorm almost everything that could possibly go wrong and all the adverse events in the environment that might affect the system. The system could then be designed in such a way as to avoid those issues. Traditional software systems also allow for formal verification guarantees. While these guarantees don't cover all possible applications, traditional software (including older types of AI and ML systems) are sufficiently understandable to humans that experts are able to analyze them.

Following design and verification, the system could be tested in situations that closely mimic the situations it would encounter in practice, and any problem that the tests uncovered could be addressed. If a problem could not be adequately addressed, then often it was possible to further constrain the system's operating environment or domain of application so that it would not be used in situations that might result in harm. This process of brainstorming what could go wrong, verification, and testing might be very resource intensive and require many iterations to get right. However, it can be done with enough success that users can be fairly confident that the final product will be within the range of what regulators deem to be acceptable risk.

Modern AI systems, however, have proven to be much more difficult to assure. Numerous aspects of modern systems contribute to this increased difficulty:

- **Complex environments and applications:** As AI increases in capabilities, it is applied to new and much more complex environments. For example, while aircraft have had autopilot systems for many years, these have been designed to work only in the relatively constrained environments of airport runways and the open air. Now, however, AI is being applied to self-driving cars, which operate in the much more complex environments of busy cities, involving numerous other cars and obstacles and a myriad of unusual driving situations. Because of these more complex environments and applications, it is much harder to anticipate all the things that might go wrong or all the adverse situations the system might encounter. In many cases the difficulty of brainstorming what could go wrong becomes great enough that it is practically impossible to do completely.
- **Very large solution space:** Modern ML algorithms are powerful optimizers that can successfully search a very large solution space to meet whatever objective they are given. Many of these algorithms discover solutions that completely surprise the designers. For example, in one of the famous 2016 matches between AlphaGo and the world champion Go player Lee Sedol, the AI made a move that all of the professional observers thought must have been a coding error, but which ended up leading to AlphaGo winning the match. AI algorithms discovering very surprising solutions is not

such a new phenomenon.[5] However, the greatly increased optimization power of recent algorithms has made this problem much more acute.

- **Extremely complex learned models:** The models learned by earlier ML algorithms were relatively simple. For example, linear regression attempts to map the relationship between just two variables using a simple linear equation. Newer algorithms, however, are learning incredibly complex models that no human could hope to completely understand within a reasonable time, or even at all. This makes all the other challenging aspects of modern AI even more difficult, because we don't really understand what the AI is doing so as to anticipate what it might do in real-world settings.
- **Online learning:** The problem of anticipating what an AI system would do under different conditions is made much more difficult when dealing with ML systems that continue to update their own models after deployment (online learning).
- **Emergent behaviors:** The challenges presented by these factors are made even more difficult when AI systems interact with other AIs or with humans. It is extraordinarily difficult to anticipate the emergent effects that can result when multiple highly-capable AIs interact with each and with humans in complex environments.

Challenge Examples

These factors lead to many new challenges. The following paragraphs give brief descriptions of some of these.

Specification challenges: Specification challenges arise when trying to ensure that an AI system's behavior aligns with the operator's true intentions.[12] Every optimization algorithm requires one or more objective functions for which it is trying to optimize in some way. However, it is often difficult or intractable to accurately capture what we truly want the system to optimize for. Instead, designers often use proxy objectives that are easier to specify. But this can lead to problems. When a powerful optimizer is given an objective function with a large possible solution space, the optimizer will often find extremely novel solutions that were completely unanticipated by the designers. If there is any gap whatsoever between what the operator actually wants versus what the designer specifies in practice, then the optimizer is liable to find very novel and potentially dangerous solutions within that gap. Several examples of such behavior have been observed in modern AI systems.[13],[14]

Approaches to goal specification issues mostly focus on the idea that the AI should try to infer the operator's preferences rather than have the operator try to explicitly specify those preferences.[15] Particular approaches to learning human preferences (or value learning) include different forms of inverse reinforcement learning[16] and imitation learning,[17] along with several other ideas.[18]

Out of distribution (OOD) robustness: Modern AI / ML systems are often very brittle and don't generalize very well to new situations. However, as AI systems are being used in increasingly complex situations, it becomes almost inevitable that these systems will encounter new

[5] See for example the evolved radio [11].

situations. These situations are “out of distribution,” meaning they do not come from the same distribution of environments that was used to train the system.[6]

There are several reasons why modern AI systems are so brittle. First, ML can only learn from the data it uses for training. If the training data does not adequately represent the situations that the AI will encounter after deployment, then the AI will likely not be able to generalize appropriately. Because of the complexity of the environment, it is difficult or impossible to adequately represent all eventualities in the training data. Additionally, the complexity and black-box nature of the learned models makes it difficult or impossible for humans to assess why a particular model is not generalizing properly.

This challenge is well-known in the ML research community and has received a lot of attention. There are many active research directions on this problem. Some of the most important include uncertainty estimation (trying to accurately reflect how confident the AI is in its assessments, so that it can know when it is unsure about something), transfer learning (training a ML algorithm on one task and then adapting some of the learned information for a different task), and meta-learning (learning better learning strategies). Other research directions include anomaly detection, robustness certification, and domain adaptation, among others. (See for example, [19], [20].)

One particular subtopic within this discussion which has received significant attention is adversarial examples. These are cases where an adversary tries to trick an AI into making mistakes. Research has shown that modern AI systems are very susceptible to being tricked in predictable and easily reproducible ways. While originally discussed in the context of security against adversaries, researchers have also shown that the natural environment can create similar results as well.[21][7]

Interpretability: Many of the challenges related to AI safety and assurance are greatly exacerbated by the fact that powerful modern AI systems are essentially black boxes, where it is extremely difficult or even impossible for humans to fully understand how the AI arrives at its solutions and decisions. A large amount of current ML research is directed at devising at least partial solutions to this challenge. Some approaches focus on ways to understand what specific parts of neural networks are doing, while others attempt to build interpretable models by design, or to build interactive visualization tools that allow humans to explore what the AI is “thinking”. The difficulty of interpretability is also compounded by the complexity of trying to pin down what degree or type of interpretability is needed for different types of models or for different applications.[22]

AI security: The increasing complexity of ML models and the reliance on large amounts of training data (potentially from diverse sources) have led to unique vulnerabilities that pose great threats to the security of AI systems. The use of online learning techniques further exacerbates these issues. As mentioned in regard to out of distribution robustness, research on adversarial

[6] The problem of OOD robustness is also known as generalization, domain shift, or distributional shift.

[7] See also the section on AI security below.

vulnerabilities and defenses for artificial neural networks has been a high-profile topic of consistent interest since the discovery and characterization of the issue in 2013.[23] Additional challenges relate to generating and defending against inference time attacks (e.g. adversarial patches), training time attacks (e.g. Trojans[24], [25]), and generating false or synthetic media (e.g. “deepfakes”[26]), among other challenges. Initial investigations into online data poisoning attacks[27],[28] have begun to look into security concerns resulting from online learning and adaptation. Of particular interest from the perspective of ML-enabled autonomous systems are the research into the impact of adversarial vulnerabilities in real-world environments (see for example a recent ECCV workshop focused on computer vision applications[29]).

Human-machine interactions: As AI-enabled autonomous systems are increasingly applied to assist or perform complex cognitive tasks, unique safety challenges will arise as a consequence of the resulting complex human-machine interactions (HMI) for safety-critical applications. The envisioned interactions between human and AI agents are expected to be more dynamic, less structured, and often stochastic in nature.

Aside from ensuring the physical safety and social comfort (“perceived safety”) of humans, there are numerous remaining open challenges in HMI related to AI safety and assurance. For example, AIs need to be able to infer or predict the actions and intentions of the humans they are interacting with. For many applications there are issues of smooth and reliable hand-off from machines to humans or vice versa - especially given known human psychological weaknesses in this area. There are also issues related to ensuring that humans do not come to trust AI systems where they are unreliable, or conversely to not trust AI systems where they are in fact reliable.[30]–[33]

Systems and Safety Engineering: Systems engineering is an interdisciplinary field that focuses on design, integration, and management of complex systems over their lifecycles. Safety engineering further emphasizes the identification of safety hazards and prediction of their potential impact, with the ultimate goal of minimizing risks and severity of impacts.[34] Over the past few decades, safety engineering has developed a rich set of analysis methodologies and tools (e.g., probabilistic risk assessment, failure mode analysis, fault-tree analysis, and reliability engineering) to ensure the safety of complex systems across safety-critical domains such as energy infrastructure and aviation industry.

Traditional systems and safety engineering is still extremely valuable as an organizing principle when considering safety and assurance of advanced ML-enabled systems. However, significant revision and generalization is necessary to accommodate unique aspects of ML development and applications, such as online learning, the heavy dependency on training data, and the greatly increased complexity and uncertainty involved in new system models and applications. Recent efforts from the autonomous systems research community have begun to address these challenges, for example [35], [36]. See also [37].

Testing, Evaluation, Verification, and Validation (TEVV): Verification refers to determining that a system meets operational and design requirements, while validation involves showing that the system performs as expected and meets the needs of users. Goals for verification often relate to safe and suitable operations. Such safety requirements become restrictions on the behavior of the autonomous system, so the adherence to these requirements must be verified. Validation testing is also needed in order to understand if the performance of an autonomous system meets user expectations.

Traditionally, TEVV requires that engineers and designers sit down with domain experts and discuss all of the things that could plausibly go wrong within the constraints of the operating environment. As mentioned earlier, however, the increased complexity and black-box nature of modern AI systems, combined with the increased complexity of the operating environment, makes this brainstorming process extremely difficult or infeasible.

Some approaches to this issue involve work on red teaming, adversarial testing, improving audit trails, and various institutional verification mechanisms.[22] There has also been some recent progress in formal verification techniques for neural networks (e.g., [38], [39]). Progress on many of the other challenges mentioned above would help with TEVV as well, especially progress on AI interpretability.

Governance and regulation: Beyond all of the technical engineering challenges mentioned above, there are also many challenges related to the governance of AI. Governing bodies need to determine the correct guidelines, regulations, and specific requirements needed to ensure the safety, security, and ethical implementation of powerful AI applications. There need to be formal ways to verify that applications and organizations are performing as required, and there needs to be legal standards worked out for who should be held accountable when something goes wrong. Additionally, it will likely be necessary to create mechanisms and institutions that prevent stakeholders from cutting corners or reneging on commitments so as to maintain their edge against competitors or adversaries.

Longer-term concerns: In the longer term, very advanced AI is likely to increase the difficulty of all of the above challenges, as well as the consequences of mistakes. For example, specification issues for very advanced AI systems may require encoding all human preferences - a nearly impossible task.[8] Governmental challenges may also require new forms of international governance and cooperation.

Additionally, there are some challenges that may only become salient with very advanced longer-term AI systems. For example, control issues become critical if highly autonomous agents are ever made capable enough that they may be able to “outsmart” their designers.[40] Some AI researchers have also argued that the theoretical foundations of decision theory, game theory, and AI theory will need to be reworked in order to accommodate autonomous agents that learn about themselves as well as the environment (“embedded agency”[41]) and to accommodate optimization agents that may themselves be capable of creating new optimizers (“mesa-optimization”[42]).

[8] This is known as the Alignment Problem.

Related Work

There are many existing literature reviews that address AI safety and assurance challenges from different perspectives. From the perspective of safety and systems engineering, for example, there are [35]–[37], [51]. From the technical machine learning perspective there is [15], [52]–[54]. And from the longer-term risk perspective there is [55], [56]. There are also numerous reviews that address particular challenges or for particular applications, many of which have been cited above in the relevant sections.

To our knowledge, however, no other report aims to be a comprehensive review of the entire field of AI safety & assurance as this report does. Additionally, to our knowledge no other report aims to incorporate perspectives and topics from all of the relevant disciplines.

There are also several thorough recommendation reports for policy makers, for example [57]. However, those reports are specifically meant to provide a succinct review of the issues for the purpose of policy decision making, with a focus on making specific policy recommendations. By contrast, this report aims to be useful both for engineers and for policy makers who want to get a deeper understanding of the issues.

One other international effort that is closely related to this report is the CLAIS Knowledge Graph System (CKGS) project from the Consortium on the Landscape of Artificial Intelligence Safety (CLAIS). Similar to this report, the CKGS is meant to be a comprehensive view of the fields of AI safety and assurance, and is meant to address multiple perspectives. This report is meant to dive deeper into the challenges and solutions than the knowledge graph, but ultimately the two projects are meant to complement each other.

Overview of Sections

[The section outline for the rest of this report is still tentative. See here for a more detailed tentative outline and a description of the intended scope.]

Representative Use Cases

In Part V of this report we will focus in on the following three representative use cases as illustrations and concrete applications for the issues and challenges discussed in the rest of this report:

1) Health informatics: AI systems have long been used in healthcare. At present they are being used for radiology, diagnosis, natural language processing for electronic health records, drug discovery, and the potential for “personalized medicine,” among other applications.[9] For the most part these systems are dedicated computer programs designed for one task at a time, mostly related to passive perception and prediction tasks. This makes such AI programs ideal

[9] For a recent overview of AI applications in healthcare, see [43]. For a survey on cutting-edge applications of deep learning in healthcare, see [44].

for considering safety and assurance challenges related to passive AI systems used in safety-critical applications. (For example, misdiagnosis or false negatives in image recognition software.)

2) Self-driving cars: While limited self-driving abilities have been available for several years, the promise of fully autonomous self-driving cars is still a ways off. To a large degree the delay is due to the lack of sufficient safety, security, and assurance guarantees for such vehicles. Self-driving cars may be excellent at driving under normal or controlled driving conditions, but they do not do so well in unusual conditions. They also present difficult challenges related to human-AI interactions and to the integration of AI into larger systems of components. The auto industry has poured an enormous amount of resources into developing autonomous vehicles and fixing the safety gaps, generating a large amount of literature in the process.[10]

3) Smart cities: Artificial intelligence and machine learning, along with the Internet of Things (IoT), have allowed cities to start incorporating “smart” applications as never before. Many cities have been investigating such applications for different aspects of urban management and design, including city planning, energy management, traffic, waste management, environmental protection, and emergency services.[50] AI and ML have also allowed these services to be intelligently integrated and managed in an autonomous or semi-autonomous manner. This makes smart cities an ideal use case for considering issues that arise when multiple AI systems interact with each other and with society as a whole.

References

[1] D. Zhang *et al.*, *The AI Index 2021 Annual Report*. AI Index Steering Committee, Human-Centered AI Institute, Stanford University, Stanford, CA, 2021.

[2] C. S. Smith, “A.I. Here, There, Everywhere,” *The New York Times*, Feb. 23, 2021. Accessed: Mar. 18, 2021. [Online]. Available: https://www.nytimes.com/2021/02/23/technology/ai-innovation-privacy-seniors-education.html

[3] “State of AI Report 2020 - ONLINE,” *Google Docs*. https://docs.google.com/presentation/d/1ZUimafgXCBSLsgbacd6-a-dqO7yLyzIl1ZJbiCBUUT4/edit?usp=sharing&usp=embed_facebook (accessed Mar. 18, 2021).

[4] U. of York, “Body of Knowledge definitions,” *University of York*. https://www.york.ac.uk/assuring-autonomy/body-of-knowledge/definitions/ (accessed Mar. 12, 2021).

[5] .

[6] K. P. Murphy, *Machine learning: a probabilistic perspective*. Cambridge, Massachusetts London, England: The MIT Press, 2012.

[7] I. Sommerville, “Critical systems,” *Software Engineering 10th edition*, Oct. 07, 2014. https://iansommerville.com/software-engineering-book/static/web/critical-systems/ (accessed Mar. 12, 2021).

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” *Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, Accessed: Mar. 12, 2021. [Online]. Available:

[10] For more on safety and assurance challenges related to AI in self-driving cars, see [45]–[49].

https://papers.nips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

[9] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. Cambridge, Massachusetts London, England: The MIT Press, 2016.

[10] T. Chivers, *The AI does not hate you: the rationalists and the race to save the world*. 2019.

[11] J. Bird and P. Layzell, “The evolved radio and its implications for modelling the evolution of novel sensors,” in *Proceedings of the 2002 Congress on Evolutionary Computation. CEC’02 (Cat. No.02TH8600)*, May 2002, vol. 2, pp. 1836–1841 vol.2. doi: 10.1109/CEC.2002.1004522.

[12] D. S. Research, “Building safe artificial intelligence: specification, robustness, and assurance,” *Medium*, Sep. 27, 2018. https://medium.com/@deepmindsafetyresearch/building-safe-artificial-intelligence-52f5f75058f1 (accessed Mar. 12, 2021).

[13] “Faulty Reward Functions in the Wild,” *OpenAI*, Dec. 22, 2016. https://openai.com/blog/faulty-reward-functions/ (accessed Mar. 12, 2021).

[14] “Specification gaming examples in AI - master list - Google Drive.” https://docs.google.com/spreadsheets/d/e/2PACX-1vRPiprOaC3HsCf5Tuum8bRfzYUiKLRqJmbOoC-32JorNdfyTiRRsR7Ea5eWtvsWzuxo8bjOxCG84dAg/pubhtml (accessed Mar. 12, 2021).

[15] S. J. Russell, *Human compatible: artificial intelligence and the problem of control*. 2019.

[16] A. Y. Ng and S. Russell, “Algorithms for Inverse Reinforcement Learning,” in *in Proc. 17th International Conf. on Machine Learning*, 2000, pp. 663–670. [Online]. Available: https://ai.stanford.edu/~ang/papers/icml00-irl.pdf

[17] A. Hussein, M. M. Gaber, E. Elyan, and C. Jayne, “Imitation Learning: A Survey of Learning Methods,” *ACM Comput. Surv.*, vol. 50, no. 2, p. 21:1-21:35, Apr. 2017, doi: 10.1145/3054912.

[18] E. Bıyık, D. P. Losey, M. Palan, N. C. Landolfi, G. Shevchuk, and D. Sadigh, “Learning Reward Functions from Diverse Sources of Human Feedback: Optimally Integrating Demonstrations and Preferences,” *ArXiv200614091 Cs*, Jun. 2020, Accessed: Mar. 18, 2021. [Online]. Available: http://arxiv.org/abs/2006.14091

[19] G. Wilson and D. J. Cook, “A Survey of Unsupervised Deep Domain Adaptation,” *ArXiv181202849 Cs Stat*, Feb. 2020, Accessed: Mar. 18, 2021. [Online]. Available: http://arxiv.org/abs/1812.02849

[20] J. M. Cohen, E. Rosenfeld, and J. Z. Kolter, “Certified Adversarial Robustness via Randomized Smoothing,” *ArXiv190202918 Cs Stat*, Jun. 2019, Accessed: Mar. 18, 2021. [Online]. Available: http://arxiv.org/abs/1902.02918

[21] D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt, and D. Song, “Natural Adversarial Examples,” *ArXiv190707174 Cs Stat*, Mar. 2021, Accessed: Mar. 18, 2021. [Online]. Available: http://arxiv.org/abs/1907.07174

[22] M. Brundage *et al.*, “Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims,” *ArXiv200407213 Cs*, Apr. 2020, Accessed: Apr. 28, 2020. [Online]. Available: http://arxiv.org/abs/2004.07213

[23] C. Szegedy *et al.*, “Intriguing properties of neural networks,” *ArXiv13126199 Cs*, Feb. 2014, Accessed: Mar. 18, 2021. [Online]. Available: http://arxiv.org/abs/1312.6199

[24] T. Gu, B. Dolan-Gavitt, and S. Garg, “BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain,” *ArXiv170806733 Cs*, Mar. 2019, Accessed: Mar. 18, 2021. [Online]. Available: http://arxiv.org/abs/1708.06733

[25] K. Karra, C. Ashcraft, and N. Fendley, “The TrojAI Software Framework: An OpenSource tool for Embedding Trojans into Deep Learning Models,” *ArXiv200307233 Cs*, Mar. 2020, Accessed: Mar. 18, 2021. [Online]. Available: http://arxiv.org/abs/2003.07233

[26] Y. Mirsky and W. Lee, “The Creation and Detection of Deepfakes: A Survey,” *ACM*

Comput. Surv., vol. 54, no. 1, p. 7:1-7:41, Jan. 2021, doi: 10.1145/3425780.

[27] X. Zhang, X. Zhu, and L. Lessard, “Online Data Poisoning Attacks,” in *Learning for Dynamics and Control*, Jul. 2020, pp. 201–210. Accessed: Mar. 18, 2021. [Online]. Available: http://proceedings.mlr.press/v120/zhang20b.html

[28] E. Perry, “Lethean Attack: An Online Data Poisoning Technique,” *ArXiv201112355 Cs*, Nov. 2020, Accessed: Mar. 18, 2021. [Online]. Available: http://arxiv.org/abs/2011.12355

[29] “ECCV 2020 Workshop on Adversarial Robustness in the Real World.” https://eccv20-adv-workshop.github.io/ (accessed Mar. 18, 2021).

[30] “Socio-Cyber-Physical Systems: Models, Opportunities, Open Challenges - Research Database, The University of York.” https://pure.york.ac.uk/portal/en/publications/sociocyberphysical-systems-models-opportunities-open-challenges(578f4a0d-a43b-4498-9a1a-bb624f97a601).html (accessed Feb. 20, 2020).

[31] M. Sujan *et al.*, “Human factors challenges for the safe use of artificial intelligence in patient care,” *BMJ Health Care Inform.*, vol. 26, no. 1, Nov. 2019, doi: 10.1136/bmjhci-2019-100081.

[32] M. Konaev, T. Huang, and H. Chahal, “Trusted Partners: Human-Machine Teaming and the Future of Military AI,” Center for Security and Emerging Technology, Feb. 2021. doi: 10.51593/20200024.

[33] A. Dafoe *et al.*, “Open Problems in Cooperative AI,” *ArXiv201208630 Cs*, Dec. 2020, Accessed: Mar. 18, 2021. [Online]. Available: http://arxiv.org/abs/2012.08630

[34] SEBoK, *Safety Engineering — SEBoK*. 2020. [Online]. Available: https://www.sebokwiki.org/w/index.php?title=Safety_Engineering&oldid=60250

[35] R. Ashmore, R. Calinescu, and C. Paterson, “Assuring the Machine Learning Lifecycle: Desiderata, Methods, and Challenges,” *ArXiv190504223 Cs Stat*, May 2019, Accessed: Feb. 20, 2020. [Online]. Available: http://arxiv.org/abs/1905.04223

[36] R. Hawkins, C. Paterson, C. Picardi, Y. Jia, R. Calinescu, and I. Habli, “Guidance on the Assurance of Machine Learning in Autonomous Systems (AMLAS),” *ArXiv210201564 Cs*, Feb. 2021, Accessed: Mar. 16, 2021. [Online]. Available: http://arxiv.org/abs/2102.01564

[37] L. Fischer *et al.*, “AI System Engineering—Key Challenges and Lessons Learned,” *Mach. Learn. Knowl. Extr.*, vol. 3, no. 1, Art. no. 1, Mar. 2021, doi: 10.3390/make3010004.

[38] I. Papusha, R. Wu, J. Brulé, Y. Kouskoulas, D. Genin, and A. Schmidt, “Incorrect by Construction: Fine Tuning Neural Networks for Guaranteed Performance on Finite Sets of Examples,” *ArXiv200801204 Cs Math Stat*, Aug. 2020, Accessed: Mar. 18, 2021. [Online]. Available: http://arxiv.org/abs/2008.01204

[39] M. Fazlyab, M. Morari, and G. J. Pappas, “Safety Verification and Robustness Analysis of Neural Networks via Quadratic Constraints and Semidefinite Programming,” *ArXiv190301287 Cs Math*, Jun. 2020, Accessed: Mar. 18, 2021. [Online]. Available: http://arxiv.org/abs/1903.01287

[40] R. V. Yampolskiy, “On Controllability of AI,” *ArXiv200804071 Cs*, Jul. 2020, Accessed: Mar. 18, 2021. [Online]. Available: http://arxiv.org/abs/2008.04071

[41] A. Demski and S. Garrabrant, “Embedded Agency,” *ArXiv190209469 Cs*, Oct. 2020, Accessed: Mar. 16, 2021. [Online]. Available: http://arxiv.org/abs/1902.09469

[42] E. Hubinger, C. van Merwijk, V. Mikulik, J. Skalse, and S. Garrabrant, “Risks from Learned Optimization in Advanced Machine Learning Systems,” *ArXiv190601820 Cs*, Jun. 2019, Accessed: Mar. 16, 2021. [Online]. Available: http://arxiv.org/abs/1906.01820

[43] G. H. Kwak and P. Hui, “DeepHealth: Review and challenges of artificial intelligence in health informatics,” *ArXiv190900384 Cs Eess Stat*, Aug. 2020, Accessed: Mar. 15, 2021. [Online]. Available: http://arxiv.org/abs/1909.00384

[44] J. Egger, C. Gsaxner, A. Pepe, and J. Li, “Medical Deep Learning – A systematic Meta-Review,” p. 54.

[45] R. Bloomfield *et al.*, “Towards Identifying and closing Gaps in Assurance of autonomous Road vehicleS -- a collection of Technical Notes Part 1,” *ArXiv200300789 Cs Eess*, Feb. 2020, Accessed: Mar. 16, 2021. [Online]. Available: http://arxiv.org/abs/2003.00789

[46] R. Bloomfield *et al.*, “Towards Identifying and closing Gaps in Assurance of autonomous Road vehicleS -- a collection of Technical Notes Part 2,” *ArXiv200300790 Cs Eess*, Feb. 2020, Accessed: Mar. 16, 2021. [Online]. Available: http://arxiv.org/abs/2003.00790

[47] N. Marko, “Challenges of engineering safe and secure highly automated vehicles”,” p. 13.

[48] X. Di and R. Shi, “A Survey on Autonomous Vehicle Control in the Era of Mixed-Autonomy: From Physics-Based to AI-Guided Driving Policy Learning,” *ArXiv200705156 Cs*, Jul. 2020, Accessed: Mar. 15, 2021. [Online]. Available: http://arxiv.org/abs/2007.05156

[49] S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu, “A Survey of Deep Learning Techniques for Autonomous Driving,” *J. Field Robot.*, vol. 37, no. 3, pp. 362–386, Apr. 2020, doi: 10.1002/rob.21918.

[50] K. Ahmad, M. Maabreh, M. Ghaly, K. Khan, J. Qadir, and A. Al-Fuqaha, “Developing Future Human-Centered Smart Cities: Critical Analysis of Smart City Security, Interpretability, and Ethical Challenges,” *ArXiv201209110 Cs*, Dec. 2020, Accessed: Mar. 15, 2021. [Online]. Available: http://arxiv.org/abs/2012.09110

[51] U. of York, “Body of Knowledge,” *University of York*. https://www.york.ac.uk/assuring-autonomy/body-of-knowledge/ (accessed Mar. 18, 2021).

[52] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, “Concrete Problems in AI Safety,” *ArXiv160606565 Cs*, Jul. 2016, Accessed: Feb. 20, 2020. [Online]. Available: http://arxiv.org/abs/1606.06565

[53] B. Christian, *The alignment problem: machine learning and human values*. 2020.

[54] J. Whittlestone, K. Arulkumaran, and M. Crosby, “The Societal Implications of Deep Reinforcement Learning,” *J. Artif. Intell. Res.*, vol. 70, pp. 1003-1030-1003–1030, Mar. 2021, doi: 10.1613/jair.1.12360.

[55] T. Everitt, G. Lea, and M. Hutter, “AGI Safety Literature Review,” *ArXiv180501109 Cs*, May 2018, Accessed: Feb. 20, 2020. [Online]. Available: http://arxiv.org/abs/1805.01109

[56] E. Hubinger, “An overview of 11 proposals for building safe advanced AI,” *ArXiv201207532 Cs*, Dec. 2020, Accessed: Mar. 18, 2021. [Online]. Available: http://arxiv.org/abs/2012.07532

[57] National Security Commission on Artificial Intelligence, “National Security Commission on Artificial Intelligence Final Report,” Jan. 2021. [Online]. Available: https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf

II. Algorithmic Challenges

1. Specification Issues

Aryeh Englander

Introduction

As mentioned in the Overview, specification challenges arise when an AI algorithm achieves its specified objectives in a way that is not aligned with the designer's true intentions. Every optimization algorithm requires one or more objective functions which the algorithm is trying to optimize in some way. However, it is often difficult or intractable for designers to fully capture their intentions in an explicit objective function. Additionally, when a powerful AI optimizer is given an objective function with a large possible solution space, the optimizer will often find extremely novel solutions that were completely unanticipated by the designers. If there is any gap between the AI's specified objective and what the designer actually wants, then the AI is liable to find unanticipated and potentially dangerous solutions within that gap.[11]

Specification challenges come in a few varieties. *Specification gaming* occurs when the algorithm technically achieves the objective as specified, but in a way that differs from the operator's true intent.[12] For example, in one case researchers trained a reinforcement learning agent to play a boat racing game by trying to maximize the agent's score. While human players naturally try to get the most points primarily by winning the game, the game also awarded points for hitting certain waypoints. The AI discovered that in one particular location it could enter an infinite loop of hitting waypoints, thereby maximizing its score without ever ending the race. In retrospect, the proxy goal of maximizing score was flawed, but this was not at all obvious beforehand.[4] Numerous other examples of specification gaming have also been observed in modern AI systems.[5]

Another type of specification issue occurs when the algorithm does achieve the intended objective, but does so in a way that causes unintended *negative side effects* along the way. For example, an operator may assign a robot the task of moving from point A to point B may, but forget to specify that the robot should not knock over anything along the way.[6]

Specification challenges are not limited to reinforcement learning. They have also been repeatedly observed in evolutionary algorithms [7], and there has been discussion of how they might arise in modern language models [8]. Some AI bias and fairness issues can also be

[11] Specification challenges are closely related to the challenge of intent alignment - i.e., ensuring that the algorithm's intent is aligned with that of the designers.[1] Specification challenges in AI are also related to a broader phenomenon known from statistics and the social sciences, often referred to as Goodhart's Law: "When a measure becomes a target, it ceases to be a good measure."[2] See [3] for an analysis of Goodhart-type phenomena and how they relate to AI.

[12] In the reinforcement learning context, specification gaming is also often referred to as reward gaming or reward hacking.

viewed as a type of specification gaming. For example, in 2015 a Google image recognition algorithm mistakenly categorized some black people as gorillas. Part of the problem was that the designers specified that the algorithm should optimize for *average* accuracy across all image categories, rather than specifying that the algorithm should prioritize the accuracy of certain categories (e.g., humans) over others.[9]

The types of problems caused by misspecified objectives can also occur due to training data that does not fully capture what the designers intend the algorithm to learn, or from baked-in assumptions, hyperparameters, or other algorithmic details that end up leading the algorithm to solutions that are not aligned with the type of solutions that the designers intended.[5], [8] Many of these issues are partially covered in other sections of this document. See especially the sections on out-of-distribution robustness, data management, and bias and fairness.

Existing Approaches

The traditional approach to misspecification is to use testing, evaluation, verification, and validation (TEVV) techniques to determine if and when an algorithm is safe. If the algorithm is found to be safe only under certain operating conditions, then the algorithm is regulated so as to restrict its use to situations where those operating conditions are met. Monitoring and failsafe mechanisms are also often put in place as a further precaution.

However, as discussed in the Overview, the traditional approach is inadequate for many modern AI-enabled systems. The increase in AI capabilities and complexity often makes it difficult or impossible to anticipate possible failure modes, especially for systems that continue to learn after deployment or which interact with humans or other AI-enabled systems. Additionally, the increased complexity of the environments in which these systems are deployed often makes it infeasible to test a sufficiently broad range of scenarios that the system is likely to encounter.

Because of these issues, some researchers have started to shift away from directly specifying the algorithm's objectives, and to instead have the algorithm infer the preferences of its human operators from data.

One straightforward approach to learning human preferences is known as *imitation learning*, where the AI tries to mimic the way that humans approach the task. Imitation learning, however, is limited by the inability of humans to provide sufficient demonstrations to cover all scenarios that the agent might encounter after deployment. Imitation also cannot generally allow the agent to improve beyond human performance levels.[10] (See [11] for a survey of imitation learning.)

Another approach is *reward learning*, where the AI infers the objective (rather than a particular policy or set of rules) by observing a dataset of relevant human choices. *Inverse reinforcement learning (IRL)*, perhaps the most popular goal learning approach, uses demonstrations from humans which implicitly contain a set of choices. Other datasets for observing human choices include preference comparisons [12], proxy rewards [13], and natural language [14], among others. (See [15]–[19] for more on IRL and other reward learning approaches.)

Assistance games (or *cooperative inverse reinforcement learning* [20]) are a variation on reward learning where the human simultaneously helps the AI learn the objective and accomplish the task. Standard reward learning, by contrast, first learns the objective and only afterwards tries (without human assistance) to achieve that objective. (See [18] for an analysis of assistance games vs. reward learning.)

Rather than focus on learning human preferences, some approaches to specification issues focus instead on trying to avoid negative side effects, for example by penalizing impacts on the environment [21] or on the ability to reach other goals [22]. See also [23]**.**

Open Challenges

One problem with goal learning and assistance games is that they typically assume that the observed humans accurately know the true objective and are optimizing towards it. This is questionable given what we currently know about human cognition. [17], [18] identify additional challenges facing the goal learning and assistance games approaches.

Most current research has focused on reinforcement learning algorithms and on interactions between individual AI agents and individual humans. Comparatively little has been written about specification gaming in other types of AI algorithms, such as language models, and in multi-agent or human-machine teaming scenarios. (For some research on these topics, see [8], [24]–[28].)

A sufficiently advanced AI system may learn to run its own optimization processes in pursuit of achieving its overall objectives. Ensuring that learned optimization processes of this type remain aligned is an even more difficult challenge than trying to align the AI’s overall objectives.[29]

Very advanced AI systems may also be capable of deliberately trying to fool their operators into thinking that they are aligned when in fact they are not, or to interfere with attempts by human operators to correct issues after deployment. AI systems that actively try to avoid human interference are known as *incorrigible* systems, and preventing such behavior is a particularly difficult challenge.[30] If a sufficiently advanced AI system has the ability to modify its own reward-generating mechanism, it may even try to “hack” the system to generate rewards regardless of its own actions.[31], [32]

Several approaches have been proposed for dealing with specification challenges in very advanced AI.[33] Notable research programs in this direction include iterated amplification and debate [34], [35] and recursive reward modeling [16]. However, this work is still in its infancy and much more research will be required to address the issues.

References

[1] P. Christiano, "Clarifying 'AI Alignment' - AI Alignment Forum," *AI Alignment Forum*, Nov. 05, 2018. https://www.alignmentforum.org/posts/ZeE7EKHTFMBs8eMxn/clarifying-ai-alignment (accessed May 18, 2021).

[2] M. Strathern, "'Improving ratings': audit in the British University system | European Review | Cambridge Core," Jul. 13, 2009. https://www.cambridge.org/core/journals/european-review/article/abs/improving-ratings-audit-in-the-british-university-system/FC2EE640C0C44E3DB87C29FB666E9AAB (accessed May 18, 2021).

[3] D. Manheim and S. Garrabrant, "Categorizing Variants of Goodhart's Law," *ArXiv180304585 Cs Q-Fin Stat*, Feb. 2019, Accessed: May 18, 2021. [Online]. Available: http://arxiv.org/abs/1803.04585

[4] J. Clark and D. Amodei, "Faulty Reward Functions in the Wild," *OpenAI*, Dec. 22, 2016. https://openai.com/blog/faulty-reward-functions/ (accessed May 14, 2021).

[5] V. Krakovna *et al.*, "Specification gaming: the flip side of AI ingenuity," *Deepmind*, 2020. deepmind.com/blog/article/Specification-gaming-the-flip-side-of-AI-ingenuity (accessed May 14, 2021).

[6] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete Problems in AI Safety," *ArXiv160606565 Cs*, Jul. 2016, Accessed: Feb. 20, 2020. [Online]. Available: http://arxiv.org/abs/1606.06565

[7] DeepMind, "Specification gaming examples in AI - master list - Google Drive." https://docs.google.com/spreadsheets/d/e/2PACX-1vRPiprOaC3HsCf5Tuum8bRfzYUiKLRqJmbOoC-32JorNdfyTiRRsR7Ea5eWtvsWzuxo8bjOxCG84dAg/pubhtml (accessed May 14, 2021).

[8] Z. Kenton, T. Everitt, L. Weidinger, I. Gabriel, V. Mikulik, and G. Irving, "Alignment of Language Agents," *ArXiv210314659 Cs*, Mar. 2021, Accessed: Apr. 22, 2021. [Online]. Available: http://arxiv.org/abs/2103.14659

[9] S. J. Russell, *Human compatible: artificial intelligence and the problem of control*. 2019.

[10] P. Christiano, "The easy goal inference problem is still hard," *Medium*, May 13, 2015. https://ai-alignment.com/the-easy-goal-inference-problem-is-still-hard-fad030e0a876 (accessed May 20, 2021).

[11] F. Torabi, G. Warnell, and P. Stone, "Recent Advances in Imitation Learning from Observation," *ArXiv190513566 Cs*, Jun. 2019, Accessed: May 18, 2021. [Online]. Available: http://arxiv.org/abs/1905.13566

[12] C. Wirth, R. Akrour, G. Neumann, and J. Fürnkranz, "A Survey of Preference-Based Reinforcement Learning Methods," *J. Mach. Learn. Res.*, vol. 18, no. 136, pp. 1–46, 2017.

[13] D. Hadfield-Menell, S. Milli, P. Abbeel, S. Russell, and A. Dragan, "Inverse Reward Design," *ArXiv171102827 Cs*, Oct. 2020, Accessed: May 11, 2021. [Online]. Available: http://arxiv.org/abs/1711.02827

[14] J. Fu, A. Korattikara, S. Levine, and S. Guadarrama, "From Language to Goals: Inverse Reinforcement Learning for Vision-Based Instruction Following," *ArXiv190207742 Cs Stat*, Feb. 2019, Accessed: May 18, 2021. [Online]. Available: http://arxiv.org/abs/1902.07742

[15] S. Arora and P. Doshi, "A survey of inverse reinforcement learning: Challenges, methods and progress," *Artif. Intell.*, vol. 297, p. 103500, Aug. 2021, doi: 10.1016/j.artint.2021.103500.

[16] J. Leike, D. Krueger, T. Everitt, M. Martic, V. Maini, and S. Legg, "Scalable agent alignment via reward modeling: a research direction," *ArXiv181107871 Cs Stat*, Nov. 2018, Accessed: May 18, 2021. [Online]. Available: http://arxiv.org/abs/1811.07871

[17] H. J. Jeon, S. Milli, and A. D. Dragan, "Reward-rational (implicit) choice: A unifying

formalism for reward learning," *ArXiv200204833 Cs*, Dec. 2020, Accessed: May 18, 2021. [Online]. Available: http://arxiv.org/abs/2002.04833

[18] R. Shah *et al.*, "Benefits of Assistance over Reward Learning," Sep. 2020, Accessed: May 11, 2021. [Online]. Available: https://openreview.net/forum?id=DFIoGDZejIB

[19] Francis Rhys Ward, Francesca Toni, and Francesco Belardinelli, "A Survey of Reward Learning as a Solution to AI Alignment," forthcoming.

[20] D. Hadfield-Menell, A. Dragan, P. Abbeel, and S. Russell, "Cooperative Inverse Reinforcement Learning," *ArXiv160603137 Cs*, Nov. 2016, Accessed: May 11, 2021. [Online]. Available: http://arxiv.org/abs/1606.03137

[21] V. Krakovna, L. Orseau, R. Kumar, M. Martic, and S. Legg, "Penalizing side effects using stepwise relative reachability," *ArXiv180601186 Cs Stat*, Mar. 2019, Accessed: Mar. 16, 2021. [Online]. Available: http://arxiv.org/abs/1806.01186

[22] A. M. Turner, D. Hadfield-Menell, and P. Tadepalli, "Conservative Agency via Attainable Utility Preservation," *Proc. AAAIACM Conf. AI Ethics Soc.*, pp. 385–391, Feb. 2020, doi: 10.1145/3375627.3375851.

[23] D. Lindner, K. Matoba, and A. Meulemans, "Challenges for Using Impact Regularizers to Avoid Negative Side Effects," *ArXiv210112509 Cs*, Feb. 2021, Accessed: Mar. 16, 2021. [Online]. Available: http://arxiv.org/abs/2101.12509

[24] A. Ecoffet, J. Clune, and J. Lehman, "Open Questions in Creating Safe Open-ended AI: Tensions Between Control and Creativity," *ArXiv200607495 Cs*, Jun. 2020, Accessed: May 26, 2021. [Online]. Available: http://arxiv.org/abs/2006.07495

[25] J. Stray, S. Adler, and D. Hadfield-Menell, "What are you optimizing for? Aligning Recommender Systems with Human Values," p. 7.

[26] D. Manheim, "Multiparty Dynamics and Failure Modes for Machine Learning and Artificial Intelligence," *Big Data Cogn. Comput.*, vol. 3, no. 2, Art. no. 2, Jun. 2019, doi: 10.3390/bdcc3020021.

[27] A. Fickinger, S. Zhuang, D. Hadfield-Menell, and S. Russell, "Multi-Principal Assistance Games," *ArXiv200709540 Cs*, Jul. 2020, Accessed: May 26, 2021. [Online]. Available: http://arxiv.org/abs/2007.09540

[28] A. Critch and D. Krueger, "AI Research Considerations for Human Existential Safety (ARCHES)," *ArXiv200604948 Cs*, May 2020, Accessed: May 26, 2021. [Online]. Available: http://arxiv.org/abs/2006.04948

[29] E. Hubinger, C. van Merwijk, V. Mikulik, J. Skalse, and S. Garrabrant, "Risks from Learned Optimization in Advanced Machine Learning Systems," *ArXiv190601820 Cs*, Jun. 2019, Accessed: Mar. 16, 2021. [Online]. Available: http://arxiv.org/abs/1906.01820

[30] N. Soares, B. Fallenstein, and E. Yudkowsky, "Corrigibility," *AAAI Workshop AI Ethics*, 2015, [Online]. Available: https://intelligence.org/files/Corrigibility.pdf

[31] J. Uesato, R. Kumar, V. Krakovna, T. Everitt, R. Ngo, and S. Legg, "Avoiding Tampering Incentives in Deep RL via Decoupled Approval," *ArXiv201108827 Cs*, Nov. 2020, Accessed: May 27, 2021. [Online]. Available: http://arxiv.org/abs/2011.08827

[32] T. Everitt, M. Hutter, R. Kumar, and V. Krakovna, "Reward Tampering Problems and Solutions in Reinforcement Learning: A Causal Influence Diagram Perspective," *ArXiv190804734 Cs*, Mar. 2021, Accessed: May 18, 2021. [Online]. Available: http://arxiv.org/abs/1908.04734

[33] E. Hubinger, "An overview of 11 proposals for building safe advanced AI," *ArXiv201207532 Cs*, Dec. 2020, Accessed: Mar. 18, 2021. [Online]. Available: http://arxiv.org/abs/2012.07532

[34] P. Christiano, B. Shlegeris, and D. Amodei, "Supervising strong learners by amplifying weak experts," *ArXiv181008575 Cs Stat*, Oct. 2018, Accessed: May 18, 2021. [Online]. Available: http://arxiv.org/abs/1810.08575

[35] G. Irving, P. Christiano, and D. Amodei, "AI safety via debate," *ArXiv180500899 Cs Stat*, Oct. 2018, Accessed: May 18, 2021. [Online]. Available: http://arxiv.org/abs/1805.00899

[Note: The section outline for the rest of this report is still tentative. See here for a more detailed tentative outline and a description of the intended scope.]

2. Out of Distribution Robustness

3. Safe exploration

4. AI security and privacy

5. Interpretability

6. Bias and Fairness

7. Human-Machine Interactions

8. Multi-Agent Challenges

III. System Lifecycle and Engineering

1. AI Systems and Safety Engineering

2. Data Management

3. Verification and Validation

4. Testing and Evaluation

5. Runtime Monitoring and Enforcement

6. Process Assurance and Certification

IV. Ethics, Society, and Governance

1. The Human-Centric Perspective

2. Ethics

3. Societal Implications

4. Governance

5. Longer-Term Risks and Challenges

V. Sample Domains of Application

1. Health Informatics

2. Self Driving Cars

3. Smart Cities

References

Appendices

Other Landscapes and Roadmaps

Resources / Tools