AI Alignment: A Comprehensive Survey

Jiaming Ji[*,1] **Tianyi Qiu**[*,1] **Boyuan Chen**[*,1] **Borong Zhang**[*,1] **Hantao Lou**[1] **Kaile Wang**[1]
Yawen Duan[2] **Zhonghao He**[2] **Jiayi Zhou**[1] **Zhaowei Zhang**[1] **Fanzhi Zeng**[1] **Juntao Dai**[1]
Xuehai Pan[1] **Kwan Yee Ng** **Aidan O'Gara**[5] **Hua Xu**[1] **Brian Tse** **Jie Fu**[4] **Stephen McAleer**[3]
Yaodong Yang[1, ✉] **Yizhou Wang**[1] **Song-Chun Zhu**[1] **Yike Guo**[4] **Wen Gao**[1]

[1]Peking University [2]University of Cambridge [3]Carnegie Mellon University
[4]Hong Kong University of Science and Technology [5]University of Southern California

Abstract

AI alignment aims to make AI systems behave in line with human intentions and values. As AI systems grow more capable, so do risks from misalignment. To provide a comprehensive and up-to-date overview of the alignment field, in this survey, we delve into the core concepts, methodology, and practice of alignment. First, we identify four principles as the key objectives of AI alignment: Robustness, Interpretability, Controllability, and Ethicality (**RICE**). Guided by these four principles, we outline the landscape of current alignment research and decompose them into two key components: **forward alignment** and **backward alignment**. The former aims to make AI systems aligned via alignment training, while the latter aims to gain evidence about the systems' alignment and govern them appropriately to avoid exacerbating misalignment risks. On forward alignment, we discuss techniques for learning from feedback and learning under distribution shift. Specifically, we survey traditional preference modeling methods and reinforcement learning from human feedback, and further discuss potential frameworks to reach scalable oversight for tasks where effective human oversight is hard to obtain. Within learning under distribution shift, we also cover data distribution interventions such as adversarial training that help expand the distribution of training data, and algorithmic interventions to combat goal misgeneralization. On backward alignment, we discuss assurance techniques and governance practices. Specifically, we survey assurance methods of AI systems throughout their lifecycle, covering safety evaluation, interpretability, and human value compliance. We discuss current and prospective governance practices adopted by governments, industry actors, and other third parties, aimed at managing existing and future AI risks.

This survey aims to provide a comprehensive yet beginner-friendly review of alignment research topics. Based on this, we also release and continually update the website `www.alignmentsurvey.com` which features tutorials, collections of papers, blog posts, and other resources.

---

* Equal contribution.
✉ Corresponding author. Contact `<pku.alignment@gmail.com>`.
˘ Version: v4 (updated on Feb 27, 2024). The content of the survey will be continually updated.

Contents

1 Introduction

Recent advancements have seen the increasing application of capable AI systems in complex domains. For instance, Large Language Models (LLMs) have exhibited improved capabilities in multi-step reasoning (Wei et al., 2022; Wang et al., 2023c) and cross-task generalization (Brown et al., 2020b; Askell et al., 2021) in real-world deployment settings, and these abilities are strengthened with increased training time, training data, and parameter size (Kaplan et al., 2020; Srivastava et al., 2023; Hoffmann et al., 2022). The utilization of Deep Reinforcement Learning (DRL) for the control of nuclear fusion (Degrave et al., 2022) is another notable example. The increasing capabilities and deployment in high-stakes domains come with heightened risks. Various undesirable behaviors exhibited by advanced AI systems (*e.g.*, manipulation (Perez et al., 2023; Carroll et al., 2023; Sharma et al., 2024) and deception (Park et al., 2023b)) have raised concerns about the hazards from AI systems.

Consequently, these concerns have catalyzed research efforts in *AI alignment* (Soares and Fallenstein, 2014; Christian, 2020; Hendrycks et al., 2021b). AI alignment aims to make AI systems behave in line with human intentions and values (Leike et al., 2018), focusing more on the objectives of AI systems than their capabilities. Failures of alignment (*i.e.*, misalignment) are among the most salient causes of potential harm from AI. Mechanisms underlying these failures include *reward hacking* (Pan et al., 2021) and *goal misgeneralization* (Di Langosco et al., 2022), which are further amplified by *double edge components* such as situational awareness (Cotra, 2022), broadly-scoped goals (Ngo et al., 2024), mesa-optimization objectives (Hubinger et al., 2019c), and access to increased resources (Shevlane et al., 2023) (§1.1.2).

Alignment efforts to address these failures focus on accomplishing four key objectives (§1.2.2): Robustness, Interpretability, Controllability, and Ethicality (**RICE**). Current research and practice on alignment consist of four areas (§1.2): Learning from Feedback (§2), Learning under Distributional Shift (§3), Assurance (§4), and Governance (§5). The four areas and the RICE objectives are not in one-to-one correspondence. Each individual area often serves more than one alignment objective, and vice versa (see Table 1).

In this survey, we introduce the concept, methodology, and practice of AI alignment and discuss its potential future directions.[1]

1.1 The Motivation for Alignment

The motivation for alignment is a three-step argument, each step building upon the previous one: (1) Deep learning-based systems (or applications) have an increasingly large impact on society and bring significant risks ; (2) Misalignment represents a significant source of risks; and (3) Alignment research and practice address risks stemming from misaligned systems (*e.g.*, power-seeking behaviors).

1.1.1 Risks of Misalignment

With improved capabilities of AI systems, come increased risks.[2] Some undesirable behaviors of LLMs including (but not limited to) untruthful answers (Bang et al., 2023), sycophancy (Perez et al., 2023; Sharma et al., 2024), and deception (Jacob Steinhardt, 2023; Park et al., 2023b) worsen with increased model scale (Perez et al., 2023), resulting in concerns about advanced AI systems that are hard to control. Moreover, emerging trends such as *LLM-based agents* (Xi et al., 2023; Wang et al., 2023b) also raise concerns about the system's controllability and ethicality (Chan et al., 2023). Looking further ahead, the development of increasingly competent AI systems opens up the possibility of realizing Artificial General Intelligence (AGI) in the foreseeable future, *i.e.*, systems can match or surpass human intelligence in all relevant aspects (Bubeck et al., 2023). This could bring extensive opportunities (Manyika et al., 2017), *e.g.*, automation (West, 2018), efficiency improvements (Furman and Seamans, 2019), but also come with serious risks (CAIS, 2023; Critch and Russell, 2023), such as safety concerns (Hendrycks and Mazeika, 2022), biases and inequalities (Ntoutsi et al., 2020), and large-scale risks from superhuman capabilities (Bengio, 2023). Taking biases as an example, cutting-edge LLMs manifest discernible biases about gender, sexual identity, and immigrant status among others (Perez et al., 2023), which could reinforce existing inequalities.

Within the large-scale risks from superhuman capabilities, it has been conjectured that global catastrophic risks (*i.e.*, risks of severe harms on a global scale) (Bostrom and Cirkovic, 2011; Hendrycks et al., 2023; Government of the United Kingdom, 2023) and existential risks (*i.e.*, risks that threaten the destruction of humanity's long-term potential) from advanced AI systems are especially worrying. These concerns are elaborated in first-principle deductive arguments (Ngo, 2020a; Bengio, 2023), evolutionary analysis (Hendrycks, 2023), and concrete scenario mapping (Christiano, 2019; Kenton et al., 2022). In CAIS (2023), leading AI scientists and other notable figures stated that *Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war*. The median researcher surveyed by Stein-Perlman et al. (2022) at NeurIPS 2021 and ICML 2021 reported a 5% chance that the long-run effect of advanced AI on humanity would

---

[1]To help beginners interested in this field learn more effectively, we highlight resources about alignment techniques. More details can be found at `www.alignmentsurvey.com/resources`

[2]We discuss and taxonomize the risks that might brought by misaligned AI systems, please see §1.1.2.

be *extremely bad (e.g., human extinction)*, and 36% of NLP researchers surveyed by Michael et al. (2022) self-reported to believe that *AI could produce catastrophic outcomes in this century, on the level of all-out nuclear war*.[3] Existential risks from AI also include risks of lock-in, stagnation, and more (Bostrom, 2013; Hendrycks and Mazeika, 2022), in addition to extinction risks.[4] The UK have hosted the world's first global AI Safety Summit, gathering international governments, leading AI companies, civil society groups, and research experts. Its objectives are to: (1) assess the risks associated with AI, particularly at the cutting edge of its development; (2) explore how these risks can be mitigated through internationally coordinated efforts.[5] The summit culminated in the Bletchley Declaration (Summit, 2023), which highlighted the importance of international cooperation on AI safety. It was signed by representatives from 28 contries and the EU.

Current cutting-edge AI systems have exhibited multiple classes of undesirable or harmful behaviors that may contrast with human intentions (*e.g.*, power-seeking and manipulation) (Si et al., 2022; Pan et al., 2023a), and similar worries about more advanced systems have also been raised (Critch and Krueger, 2020; CAIS, 2023).[6] These undesirable or harmful behaviors not compliant with human intentions, known as *misalignment* of AI systems[7], can naturally occur even without misuse by malicious actors and represent a significant source of risks from AI, including safety hazards (Hendrycks et al., 2021b) and potential existential risks (Hendrycks et al., 2023).[8] These large-scale risks are significant in size due to the non-trivial likelihoods of (1) building superintelligent AI systems, (2) those AI systems pursuing large-scale goals, (3) those goals are misaligned with human intentions and values, and (4) this misalignment leads to humans losing control of humanity's future trajectory (Ngo, 2020a).

Solving the risks brought by misalignment requires the *alignment* of AI systems to ensure the objectives of the system are in accordance with human intentions and values, thereby averting unintended and unfavorable outcomes. More importantly, we expect the alignment techniques to be scaled to harder tasks and significantly advanced AI systems that are even smarter than humans. A potential solution is *Superalignment*[9], which aims to build a roughly human-level automated alignment researcher, thereby using vast amounts of compute to scale up and iteratively align safe superintelligence (OpenAI, 2023c).

1.1.2 Causes of Misalignment

In the above section, we have concluded the motivation for alignment from the perspective of the concern for AI risks and technical ethics. To offer a deeper understanding of alignment, we aim to further analyze why and how the misalignment issues occur. We will first give an overview of common failure modes, and then focus on the mechanism of feedback-induced misalignment, and finally shift our emphasis towards an examination of misaligned behaviors and dangerous capabilities. In this process, we introduce the concept of *double edge components*, which offer benefits for enhancing the capabilities of future advanced systems but also bear the potential for hazardous outcomes.

Overview of Failure Modes In order to illustrate the misalignment issue, we give an overview of alignment failure modes in this section, most of which can be categorized into *reward hacking*[10] and *goal misgeneralization*.

The learning process of RL can be deconstructed into two distinct phases: firstly, the creation of an agent primed for reward optimization, and secondly, the establishment of a reward process that furnishes the agent with appropriate reward signals. Within the framework of the Markov Reward Process (Marbach and Tsitsiklis, 2001; Puterman, 2014; Sutton and Barto, 2018), the former phase can be seen as the learning process related to the transition model (*e.g.*, model-based RL agents (Moerland et al., 2023)), or the development of specialized algorithms. The latter phase can be viewed as the construction of proxy rewards, which aim to approximate the true rewards derived from sources (*e.g.*, human preferences or environment) (Ng et al., 2000; Leike et al., 2018).

Reward Hacking: In practice, proxy rewards are often easy to optimize and measure, yet they frequently fall short of capturing the full spectrum of the actual rewards (Pan et al., 2021). This limitation is denoted as *misspecified rewards*.[11] The pursuit of optimization based on such misspecified rewards may lead to a phenomenon known

---

[3]However, survey results may hinge upon the exact wording of the questions and should be taken with caution.

[4]*Existential* and *extinction* risks are two concepts that are often mixed up. The latter is a subset of the former.

[5]Source from https://www.gov.uk/government/topical-events/ai-safety-summit-2023.

[6]See §1.1.2 for an introduction to specific misalignment challenges.

[7]Some of the misaligned behaviors are less risky (*e.g.*, the agent fails to clean the room as you want), however, some of them are dangerous for systems applied in the high-stakes environment (*e.g.*, the control of nuclear fusion (Degrave et al., 2022))

[8]It should be noted that misalignment cannot cover all sources of risks brought by Deep learning-based systems and other factors such as misuse and negligence also contribute to risks on society. See §1.2.3 for discussing AI safety beyond alignment.

[9]For more details on Superalignment, you can refer to https://openai.com/blog/introducing-superalignment.

[10]*Reward hacking* can also be broadly considered as a kind of *specification gaming*.

[11]A similar definition is reward misidentification in which scenario the reward function is only partially identifiable. For more details on reward misidentification, see *e.g.*, Tien et al. (2022); Skalse et al. (2023)

as *reward hacking*, wherein agents may appear highly proficient according to specific metrics but fall short when evaluated against human standards (Amodei et al., 2016; Everitt et al., 2017). The discrepancy between proxy rewards and true rewards often manifests as a sharp phase transition in the reward curve (Ibarz et al., 2018). Furthermore, Skalse et al. (2022) defines the hackability of rewards and provides insights into the fundamental mechanism of this phase transition, highlighting that the inappropriate simplification of the reward function can be a key factor contributing to reward hacking.

Misspecified rewards often occur due to a neglect of severe criteria for the outcomes, thus making specification too broad and potentially easily hacked (Victoria et al., 2020). More than poor reward design (Ng et al., 1999), the choice of training environment and simulator with bugs (Code Bullet, 2019) can both lead to AI systems failing to satisfy intended objectives. These problems stem from task specification, broadly defined as *specification gaming*, which refers to AI systems exploiting loopholes in the task specification without achieving intended outcomes.[12] (Victoria et al., 2020)

Reward tampering can be considered a special case of reward hacking (Everitt et al., 2021; Skalse et al., 2022), referring to AI systems corrupting the reward signals generation process (Ring and Orseau, 2011). Everitt et al. (2021) delves into the subproblems encountered by RL agents: (1) *tampering of reward function*, where the agent inappropriately interferes with the reward function itself, and (2) *tampering of reward function input*, which entails corruption within the process responsible for translating environmental states into inputs for the reward function. When the reward function is formulated through feedback from human supervisors, models can directly influence the provision of feedback (*e.g.*, AI systems intentionally generate challenging responses for humans to comprehend and judge, leading to feedback collapse) (Leike et al., 2018). Since task specification has its physical instantiation (*e.g.*, memory registers storing the reward signals), the AI systems deployed in the real world have the potential to practice manipulation behaviors, resulting in more hazardous outcomes (Victoria et al., 2020).

Goal Misgeneralization: Goal misgeneralization is another failure mode, wherein the agent actively pursues objectives distinct from the training objectives in deployment while retaining the capabilities it acquired during training (Di Langosco et al., 2022).[13] For instance, in *CoinRun* games, the agent frequently prefers reaching the end of a level, often neglecting relocated coins during testing scenarios. Di Langosco et al. (2022) draw attention to the fundamental disparity between capability generalization and goal generalization, emphasizing how the inductive biases inherent in the model and its training algorithm may inadvertently prime the model to learn a proxy objective that diverges from the intended initial objective when faced with the testing distribution. It implies that even with perfect reward specification, goal misgeneralization can occur when faced with distribution shifts (Amodei et al., 2016). It should be noted that goal misgeneralization can occur in any learning system, not limited to RL since the core feature is the pursuit of unintended goals (Shah and Varma, 2022). Moreover, it might be more dangerous if advanced AI systems escape control and leverage their capabilities to bring about undesirable states (Zhuang and Hadfield-Menell, 2020).

Feedback-Induced Misalignment   With the proliferation of advanced AI systems, the challenges related to reward hacking and goal misgeneralization have become increasingly pronounced in open-ended scenarios (Paulus et al., 2018; Knox et al., 2023). Gao et al. (2023) underscores that more capable agents tend to exploit misspecified rewards to a greater extent. While many current AI systems are primarily driven by self-supervision, it's worth noting that a substantial portion relies on feedback rewards derived from human advisors (Bai et al., 2022a), allowing us to introduce the mechanism of feedback-induced misalignment. The misalignment issues are particularly pressing in open-ended scenarios, and we can attribute them to two primary factors:

- **Limitations of Human Feedback**. During the training of LLMs, inconsistencies can arise from human data annotators (*e.g.*, the varied cultural backgrounds of these annotators can introduce implicit biases (Peng et al., 2022)) (OpenAI, 2023a). Moreover, they might even introduce biases deliberately, leading to untruthful preference data (Casper et al., 2023b). For complex tasks that are hard for humans to evaluate (*e.g.*, the value of game state), these challenges[14] become even more salient (Irving et al., 2018).

- **Limitations of Reward Modeling**. Training reward models using comparison feedback can pose significant challenges in accurately capturing human values. For example, these models may unconsciously learn suboptimal or incomplete objectives, resulting in reward hacking (Zhuang and Hadfield-Menell, 2020; Skalse et al., 2022). Meanwhile, using a single reward model may struggle to capture and specify the values of a diverse human society (Casper et al., 2023b).

---

[12]For more instances about specification gaming, please see Krakovna (2020)

[13]More discussion about Goal Misgeneralization can be found in §3.1.

[14]As AI systems are deployed into more complex tasks, these difficulties amplify, necessitating novel solutions such as *scalable oversight* (Cotra, 2018).

Figure 1: Dangerous Capabilities. Advanced AI systems would be incentivized to seek power because power will help them achieve their given objectives. Powerful AI systems might hack computer systems, manipulate humans, control and develop weaponry, and perform ethical violations while avoiding a shutdown. Original copyright belongs to wiki (wikipedia, 2023), based on which we have made further adjustments. We will further discuss these issues in §1.1.2.

Additionally, Huang et al. (2023); Andreas (2022); Kim et al. (2024) demonstrate that advanced AI systems exhibit patterns of goal pursuit and multi-step reasoning capability, which further aggravate the situation if the reward is not well-defined (Ngo et al., 2024; Yang et al., 2023).

Discussion: It can be challenging to distinguish goal misgeneralization from reward hacking in specific cases. For instance (Shah and Varma, 2022), LLMs are trained to generate *harmless, honest, and helpful* outputs, but LLMs may occasionally produce harmful outputs in detail, which seemingly receive low rewards in testing distribution (which could be seen as goal misgeneralization). However, in cases where labelers are incentivized to assign high rewards to responses deemed more helpful during the labeling process, the scenarios above[15] actually receive high rewards and represent a form of specification gaming (or reward hacking). The distinction between these two scenarios can be vague at times.

More research is needed to analyze the failure modes, gain a deeper understanding of reward hacking, and develop effective methods for detecting and mitigating goal misgeneralization to address the challenges of misaligned advanced AI systems.

Misaligned Behaviors and Outcomes   Drawing from the misalignment mechanism, optimizing for a non-robust proxy may result in misaligned behaviors, potentially leading to even more catastrophic outcomes. This section delves into a detailed exposition of specific **misaligned behaviors** (●) and introduces what we term **double edge components** (+). These components are designed to enhance the capability of AI systems in handling real-world settings but also potentially exacerbate misalignment issues. It should be noted that some of these **double edge components** (+) remain speculative. Nevertheless, it is imperative to discuss their potential impact before it is too late, as the transition from controlled to uncontrolled advanced AI systems may be just one step away (Ngo, 2020b). With increased model scale, a class of **dangerous capabilities** (*) (Shevlane et al., 2023) could also emerge. The **dangerous capabilities** (*) are concrete tasks the AI system could carry out; they may not necessarily be misaligned in themselves but are instrumental to actualizing extreme risks.

We first introduce the **double edge components** (+) and analyze how they act on AI systems. Then, we illustrate the **misaligned behaviors** (●) and **dangerous capabilities** (*) to show specific misalignment issues and provide directions for future alignment evaluation research.

+ **Situational Awareness**. AI systems may gain the ability to effectively acquire and use knowledge about its status, its position in the broader environment, its avenues for influencing this environment, and the potential reactions of the world (including humans) to its actions (Cotra, 2022). Similar behaviors have been observed

---

[15]Harmful but detailed responses

in LLMs (Jonas DeGrave, 2022; Evan Hubinger, 2023). Knowing the situation can help the model better understand human intent, finish tasks within its ability, and search for outlier help if needed. However, such knowledge also paves the way for advanced methods of reward hacking, heightened deception/manipulation skills, and an increased propensity to chase instrumental subgoals (Ngo et al., 2024). Consequently, it should be given priority when evaluating potentially hazardous capabilities in AI models, alongside eight other key competencies (Shevlane et al., 2023). A highly relevant discussion is whether language models possess *world models* (LeCun, 2022; Li et al., 2022b).

+ **Broadly-Scoped Goals**. Advanced AI systems are expected to develop objectives that span long timeframes, deal with complex tasks, and operate in open-ended settings (Ngo et al., 2024). Engaging in broadly-scoped planning can empower AI systems to generalize better on the OOD settings and serve as valuable assistants in realms such as human healthcare. However, it can also bring about the risk of encouraging manipulating behaviors (*e.g.*, AI systems may take some *bad* actions to achieve human happiness, such as persuading them to do high-pressure jobs[16] (Jacob Steinhardt, 2023)). Intuitively, one approach to mitigate this risk is to confine the optimizable objectives to short-sighted ones, such as predicting only the next word, thereby preventing over-ambitious planning, but such approaches limit systems' utility and may fail; for instance, source text data (*e.g.*, fiction) can help AI systems understand the intent and belief of the roles, and thus longer-term goal-directed behavior can be elicited (Andreas, 2022). Additionally, techniques such as RL-based fine-tuning (Christiano et al., 2017; Ouyang et al., 2022) or the application of chain-of-thought prompts (Wei et al., 2022) can enable models to adapt their acquired knowledge about planning to pave the way for broadly-scoped planning objectives (Jacob Steinhardt, 2023).

+ **Mesa-Optimization Objectives**. The learned policy may pursue inside objectives *when the learned policy itself functions as an optimizer* (*i.e.*, *mesa-optimizer*). However, this optimizer's objectives may not align with the objectives specified by the training signals, and optimization for these misaligned goals may lead to systems out of control (Hubinger et al., 2019c). Freeman et al. (2019); Wijmans et al. (2023) indicate that AI systems may possess implicit goal-directed planning and manifest emergent capabilities during the generalization phase.

+ **Access to Increased Resources**. Future AI systems may gain access to websites and engage in real-world actions, potentially yielding a more substantial impact on the world (Nakano et al., 2021). They may disseminate false information, deceive users, disrupt network security, and, in more dire scenarios, be compromised by malicious actors for ill purposes. Moreover, their increased access to data and resources can facilitate *self-proliferation*, posing existential risks (Shevlane et al., 2023).

• **Power-Seeking Behaviors**. AI systems may exhibit behaviors that attempt to gain control over resources and humans and then exert that control to achieve its assigned goal (Carlsmith, 2022). The intuitive reason why such behaviors may occur is the observation that for almost any optimization objective (*e.g.*, investment returns), the optimal policy to maximize that quantity would involve power-seeking behaviors (*e.g.*, manipulating the market), assuming the absence of solid safety and morality constraints. Omohundro (2008); Bostrom (2012) have argued that power-seeking is an *instrumental subgoal* which is instrumentally helpful for a wide range of objectives and may, therefore, be favored by AI systems. Turner et al. (2021) also proved that in MDPs that satisfy some standard assumptions, the optimal policies tend to be power-seeking. Perez et al. (2023) prompt LLMs to test their tendency to suggest power-seeking behaviors, find significant levels of such tendencies, and show that RLHF strengthens them. This also holds for other instrumental subgoals such as self-preservation (Bostrom, 2012; Shevlane et al., 2023). Another notable line of research is *side-effect avoidance*, which aims to address power-seeking behaviors by penalizing agentic systems for having too much influence over the environment. It covers RL systems (Eysenbach et al., 2018; Turner et al., 2020) and symbolic planning systems (Klassen et al., 2022).

• **Untruthful Output**. AI systems such as LLMs can produce either unintentionally or deliberately inaccurate output. Such untruthful output may diverge from established resources or lack verifiability, commonly referred to as *hallucination* (Bang et al., 2023; Zhao et al., 2023). More concerning is the phenomenon wherein LLMs may selectively provide erroneous responses to users who exhibit lower levels of education[17] (Perez et al., 2023). The behavior (also known as sycophancy) appears emergently at scale (Ajeya Cotra, 2021; Perez et al.,

---

[16]This behavior is due to models' over-optimization for broadly-scoped goals and this over-optimization is hard to perceive by humans

[17]Such behaviors bare termed *sandbagging* (Perez et al., 2023). They may have been learned from web text during pre-training, which suggests that supervised learning can also bring about deceptive behaviors if those behaviors are present in training data.

2023) and untruthful output has the potential to engender deception, especially as advanced AI systems gain greater access to online resources and websites (Jacob Steinhardt, 2023).

- **Deceptive Alignment & Manipulation**. Manipulation & Deceptive Alignment is a class of behaviors that exploit the incompetence of human evaluators or users (Hubinger et al., 2019a; Carranza et al., 2023) and even manipulate the training process through *gradient hacking* (Richard Ngo, 2022). These behaviors can potentially make detecting and addressing misaligned behaviors much harder.

  Deceptive Alignment: Misaligned AI systems may deliberately mislead their human supervisors instead of adhering to the intended task. Such deceptive behavior has already manifested in AI systems that employ evolutionary algorithms (Wilke et al., 2001; Hendrycks et al., 2021b). In these cases, agents evolved the capacity to differentiate between their evaluation and training environments. They adopted a strategic pessimistic response approach during the evaluation process, intentionally reducing their reproduction rate within a scheduling program (Lehman et al., 2020). Furthermore, AI systems may engage in intentional behaviors that superficially align with the reward signal, aiming to maximize rewards from human supervisors (Ouyang et al., 2022). It is noteworthy that current large language models occasionally generate inaccurate or suboptimal responses despite having the capacity to provide more accurate answers (Lin et al., 2022c; Chen et al., 2021). These instances of deceptive behavior present significant challenges. They undermine the ability of human advisors to offer reliable feedback (as humans cannot make sure whether the outputs of the AI models are truthful and faithful). Moreover, such deceptive behaviors can propagate false beliefs and misinformation, contaminating online information sources (Hendrycks et al., 2021b; Chen and Shu, 2024).

  Manipulation: Advanced AI systems can effectively influence individuals' beliefs, even when these beliefs are not aligned with the truth (Shevlane et al., 2023). These systems can produce deceptive or inaccurate output or even deceive human advisors to attain deceptive alignment. Such systems can even persuade individuals to take actions that may lead to hazardous outcomes (OpenAI, 2023a).

  Early-stage indications of such behaviors are present in LLMs,[18] recommender systems (where the system influences the users' preferences) (Kalimeris et al., 2021; Krueger et al., 2020; Adomavicius et al., 2022), and RL agents (where agents trained from human feedback adopt policies to trick human evaluators) (Amodei et al., 2017). Also, current LLMs already possess the capability needed for deception. In Spitale et al. (2023), it has been found that GPT-3 is super-human capable of producing convincing disinformation. Given all these early-stage indications, it is plausible that more advanced AI systems may exhibit more serious deceptive/manipulative behaviors.

- **Collectively Harmful Behaviors**. AI systems have the potential to take actions that are seemingly benign in isolation but become problematic in multi-agent or societal contexts. Classical game theory offers simplistic models for understanding these behaviors. For instance, Phelps and Russell (2023) evaluates GPT-3.5's performance in the iterated prisoner's dilemma and other social dilemmas, revealing limitations in the model's cooperative capabilities. Perolat et al. (2017) executes a parallel analysis focused on common-pool resource allocation. To mitigate such challenges, the emergent field of Cooperative AI (Dafoe et al., 2020, 2021) has been advancing as an active research frontier. However, beyond studies grounded in simplified game-theoretical frameworks, there is a pressing need for research in more realistic, socially complex settings (Singh, 2014). In these environments, agents are numerous and diverse, encompassing AI systems and human actors (Critch and Krueger, 2020). Furthermore, the complexity of these settings is amplified by the presence of unique tools for modulating AI behavior, such as social institutions and norms (Singh, 2014).[19]

- **Violation of Ethics**. Unethical behaviors in AI systems pertain to actions that counteract the common good or breach moral standards – such as those causing harm to others. These adverse behaviors often stem from omitting essential human values during the AI system's design or introducing unsuitable or obsolete values into the system (Kenward and Sinclair, 2021). Research efforts addressing these shortcomings span the domain of *machine ethics* (Yu et al., 2018; Winfield et al., 2019; Tolmeijer et al., 2020) and delve into pivotal questions, *e.g.*, *whom should AI align with?* (Santurkar et al., 2023), among other concerns.

* **Dangerous Capabilities**. Figure 1 outlines the dangerous capabilities that advanced AI systems might have. As AI systems are deployed in the real world, they may pose risks to society in many ways (*e.g.*, hack computer systems, escape containment, and even violate ethics). They may hide unwanted behaviors, fool human supervisors, and seek more resources to become more powerful. Moreover, **double edge components** (+) may intensify the danger and lead to more hazardous outcomes, even resulting in existential risks (Bostrom, 2013).

---

[18]Namely, the *untruthful output* that we discuss above.
[19]We cover cooperative AI research in §3.3.2 and §4.3.1.
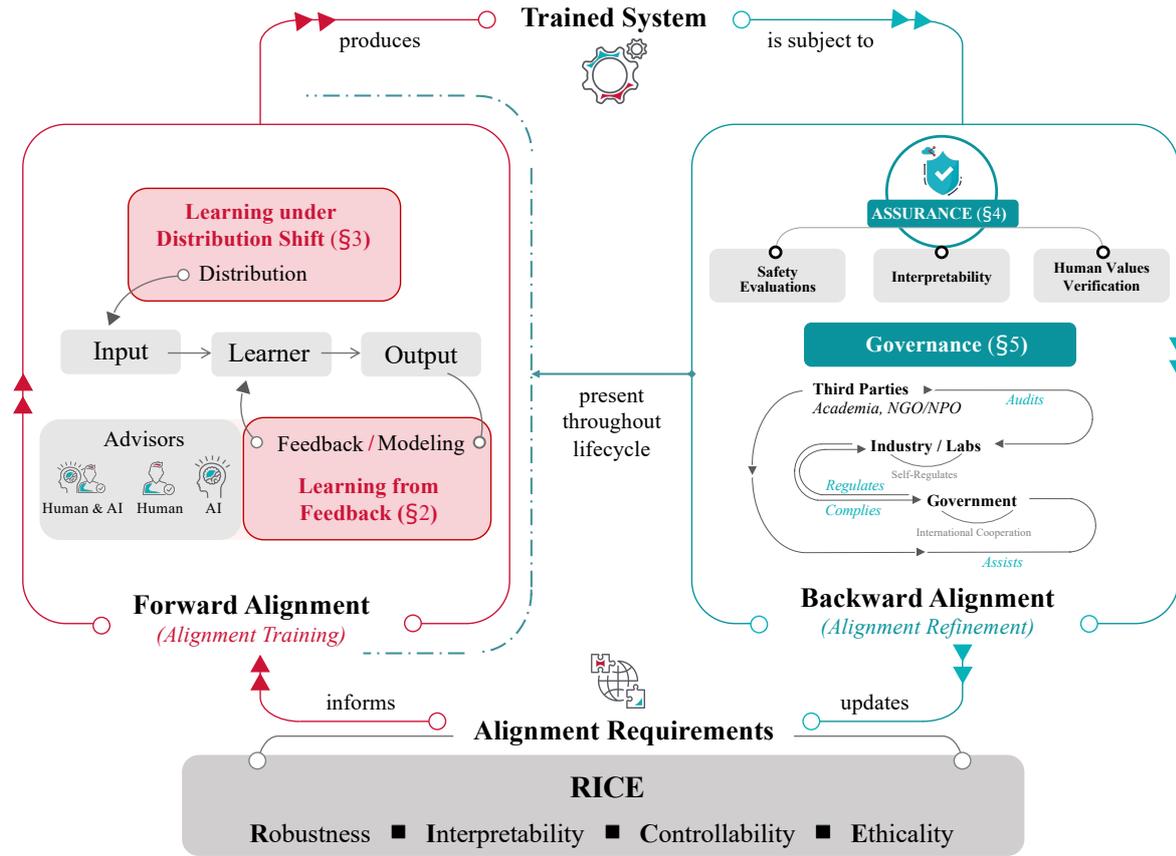
Figure 2: The Alignment Cycle. (1) **Forward Alignment** (alignment training) produces *trained systems* based on *alignment requirements*; (2) **Backward Alignment** (alignment refinement) ensures the practical alignment of *trained systems* and revises *alignment requirements*; (3) The cycle is repeated until reaching a sufficient level of alignment. Notably, although Backward Alignment has the end goal of ensuring the practical alignment of *trained systems*, it is carried out all throughout the system's lifecycle in service of this goal, including before, during, after training, and also after deployment (Shevlane et al., 2023; Koessler and Schuett, 2023; Schuett et al., 2023).

1.2   The Scope of Alignment

In this section, we focus on illustrating the scope of AI alignment: we constructed the alignment process as an *alignment cycle* and decomposed it into *Forward Alignment Process* and *Backward Alignment Process*[20] (§1.2.1). Specifically, we discuss the role of *human values* in alignment (§1.2.3) and further analyze AI safety problems beyond alignment (§1.2.3).

1.2.1   The Alignment Cycle: A Framework of Alignment

We decompose alignment into **Forward Alignment** (alignment training) (§2, §3) and **Backward Alignment** (alignment refinement) (§4, §5). Forward Alignment aims to produce trained systems that follow alignment requirements.[21] We decompose this task into Learning from Feedback (§2) and Learning under Distribution Shift (§3). Backward Alignment aims to ensure the practical alignment of the trained systems by performing evaluations in both simplistic and realistic environments and setting up regulatory guardrails to handle real-world complexities,*i.e.*, Assurance (§4). It also covers the creation and enforcement of rules that ensure the safe development and deployment of AI systems, *i.e.*, Governance (§5). At the same time, backward alignment updates the alignment requirements based on the evaluation and monitoring of the systems, both pre-deployment and post-deployment. These updated requirements then inform the next round of alignment training.

---

[20]From this point and throughout the survey, for convenience, we refer to "Forward Alignment" and "Backward Alignment".

[21]Here, *alignment requirements* refer to an operationalized specification of the alignment properties that are desired of the AI systems, including, for example, which concrete forms of robustness/interpretability/controllability/ethicality we require, in what specific settings we require them, and how they could be measured.

The two phases, forward and backward alignment, thus form a cycle where each phase produces or updates the input of the next phase (see Figure 2). This cycle, what we call *the alignment cycle*, is repeated to produce increasingly aligned AI systems. We see alignment as a dynamic process in which all standards and practices should be continually assessed and updated. Notably, Backward Alignment (including the Assurance of alignment in AI systems and the Governance of AI systems) efforts occur throughout the entire alignment cycle, as opposed to only after training. As argued in Shevlane et al. (2023); Koessler and Schuett (2023), alignment and risk evaluations should occur in every stage of the system's lifecycle, including before, during, after training, and post-deployment. Similarly, regulatory measures for every phase of the system's lifecycle have been proposed and discussed (Schuett et al., 2023; Anderljung et al., 2023).

The survey is structured around four core pillars: Learning from Feedback (§2) and Learning under Distribution Shift (§3), which constitute the components of Forward Alignment; and Assurance (§4) and Governance (§5) which form the elements of Backward Alignment. The subsequent paragraphs provide a concise introduction to each pillar, clarifying how they synergistically contribute to a comprehensive framework for AI alignment.

- **Learning from Feedback** (§2) *Learning from feedback* concerns the question of *during alignment training, how do we provide and use feedback to behaviors of the trained AI system?* It takes an input-behavior pair as given and only concerns how to provide and use feedback on this pair.[22] In the context of LLMs, a typical solution is reinforcement learning from human feedback (RLHF) (Christiano et al., 2017; Bai et al., 2022a), where human evaluators provide feedback by comparing alternative answers from the chat model, and the feedback is used via Reinforcement Learning (RL) against a trained reward model. Despite its popularity, RLHF faces many challenges (Pandey et al., 2022; Casper et al., 2023b; Tien et al., 2022), overcoming which has been a primary objective of alignment research (Bowman et al., 2022), and is one primary focus of the section. An outstanding challenge here is *scalable oversight* (§2.4), *i.e.*, providing high-quality feedback on super-human capable AI systems that operate in complex situations beyond the grasp of human evaluators, where the behaviors of AI systems may not be easily comprehended and evaluated by humans (Bowman et al., 2022). Another challenge is the problem of providing feedback on ethicality, which is approached by the direction of machine ethics (Anderson and Anderson, 2011; Tolmeijer et al., 2020). On the ethics front, misalignment could also stem from neglecting critical dimensions of variance in values, such as underrepresenting certain demographic groups in feedback data (Santurkar et al., 2023). There have also been work combining feedback mechanisms with *social choice* methods to produce a more rational and equitable aggregation of preferences (Collective Intelligence Project, 2023) (see §1.2.3).

- **Learning under Distribution Shift** (§3) In contrast to learning from feedback, which holds input fixed, this pillar focuses specifically on the cases where the distribution of input changes, *i.e.*, where distribution shift occurs (Krueger et al., 2020; Thulasidasan et al., 2021; Hendrycks et al., 2021a). More specifically, it focuses on the preservation of *alignment properties* (*i.e.*, adherence to human intentions and values) under distribution shift, as opposed to that of model capabilities. In other words, it asks how we can ensure an AI system well-aligned on the training distribution will also be well-aligned when deployed in the real world. One challenge related to distribution shift is *goal misgeneralization*, where, under the training distribution, the intended objective for the AI system (*e.g.*, following human's real intentions) is indistinguishable from other unaligned objectives (*e.g.*, gaining human approval regardless of means). The system learns the latter, which leads to unaligned behaviors in deployment distribution (Di Langosco et al., 2022). Another related challenge is *auto-induced distribution shift* (ADS), where an AI system changes its input distribution to maximize reward (Krueger et al., 2020; Perdomo et al., 2020). An example would be a recommender system shaping user preferences (Kalimeris et al., 2021; Adomavicius et al., 2022). Both goal misgeneralization and ADS are closely linked to deceptive behaviors (Park et al., 2023b) and manipulative behaviors (Shevlane et al., 2023) in AI systems, potentially serving as their causes. Interventions that address distribution shift include *algorithmic interventions* (§3.2), which changes the training process to improve reliability under other distributions, and *data distribution interventions* (§3.3) which expands the training distribution to reduce the discrepancy between training and deployment distributions. The former includes methods like Risk Extrapolation (REx) (Krueger et al., 2021) and Connectivity-based Fine-tuning (CBFT) (Lubana et al., 2023). The latter includes adversarial training (§3.3.1) (Song et al., 2018b; Bai et al., 2021) which augments training input distribution with adversarial inputs, and cooperative training (§3.3.2) (Dafoe et al., 2020, 2021) which aims to address the distribution gap between single-agent and multi-agent settings.[23]

---

[22]Here, *behavior* is broadly defined also to include the system's internal reasoning, which can be examined via interpretability tools (see §4.2).

[23]Cooperative Training aims to make AI systems more cooperative in multi-agent settings. This cooperativeness addresses multi-agent failure modes where the AI system's behavior appears benign and rational in isolation but becomes problematic within social or multi-agent scenarios (Critch and Krueger, 2020); see *collectively harmful behaviors* in §1.1.2 for a more

- **Assurance** (§4) Once an AI system has undergone forward alignment, we still need to gain confidence about its alignment before deploying it (Government of the United Kingdom, 2021; Anderljung et al., 2023). Such is the role of *assurance*: assessing the alignment of trained AI systems. Methodologies of assurance include safety evaluations (Perez et al., 2023; Shevlane et al., 2023) (§4.1) and more advanced methods such as interpretability techniques (Olah et al., 2018) (§4.2) and red teaming (Perez et al., 2022) (§4.1.3). The scope of assurance also encompasses the verification of system's alignment with human values, including formal theories focused on provable cooperativeness (Dafoe et al., 2021) and ethicality (Anderson and Anderson, 2011; Tolmeijer et al., 2020), and also a wide range of empirical and experimental methods (§4.3). Assurance takes place throughout the lifecycle of AI systems, including before, during, after training, and post-deployment, as opposed to only after training (Shevlane et al., 2023; Koessler and Schuett, 2023).[24]

- **Governance** (§5) Assurance alone cannot provide full confidence about a system's practical alignment since it does not account for real-world complexities. This necessitates governance efforts of AI systems that focus on their alignment and safety and cover the entire lifecycle of the systems (§5.1). We discuss the multi-stakeholder approach of AI governance, including the governmental regulations (Anderljung et al., 2023), the lab self-governance (Schuett et al., 2023), and the third-party practice, such as auditing (Shevlane et al., 2023; Koessler and Schuett, 2023) (§5.2). We also highlight several open problems in AI governance, including the pressing challenge of open-source governance (the governance of open-source models and the question of whether to open-source highly capable models) (Seger et al., 2023), and the importance of international coordination in AI governance (Ho et al., 2023) (§5.3). In addition to policy research, we also cover key actions from both the public and the private sector.

Comparison with Inner/Outer Decomposition   Our *alignment cycle* framework (see Figure 2) decomposes alignment into four pillars: Learning from Feedback, Learning under Distribution Shift, Assurance and Governance organized into a circular process. The design principle for this framework is three-fold: Practical (making sure pillars directly correspond to specific practices in specific stages in the system's lifecycle), Concrete (pointing to specific research directions as opposed to general themes), and Up-To-Date (accommodating and emphasizing latest developments in the alignment field). Recently, the decomposition of alignment into *outer alignment* and *inner alignment* has become popular in the alignment literature (Hubinger et al., 2019b). Outer alignment refers to the wishes of designers in accordance with the actual task specification (*e.g.*, goal & reward) used to build AI systems. And inner alignment is the consistency between task specification and the specification that the AI systems behaviors reflect (Krakovna, 2022). However, many criticisms have also been made about this characterization, including that it is ambiguous and is understood by different people to mean different things (Perry, 2020) and that it creates unnecessary difficulties by carving out problems that are not necessary conditions for success (Turner, 2022). Some have tried to remove the ambiguity by pinning down the specific causes of inner/outer misalignment and proposed, for example, *goal misspecification* and *goal misgeneralization* (Di Langosco et al., 2022; Krakovna, 2022). Learning from Feedback (approximately corresponding to *goal misspecification* and *outer alignment*) and Learning under Distribution shift (approximately corresponding to *goal misgeneralization* and *inner alignment*) in our framework tries to further improve upon the inner/outer decomposition by clarifying the exact approaches taken to address the challenges and resolving the ambiguity. Assurance and Governance, on the other hand, expands the scope to cover topics beyond outer and inner alignment.

Theoretical Research in Alignment   The alignment research literature also contains a wealth of theoretical work (Amodei et al., 2016; Everitt et al., 2018; Hendrycks et al., 2021b). These works often propose new directions and provide a foundation for practical and empirical research to build upon. We give a brief overview of this body of theoretical research below:

- **Conceptual Frameworks**. Some theoretical work proposes conceptual frameworks or characterizes subproblems within alignment. Examples include *instrumental convergence* (wherein highly intelligent agents tend to pursue a common set of sub-goals, such as self-preservation and power-seeking) (Omohundro, 2008; Bostrom, 2012), *mesa-optimization* (wherein the learned ML model performs optimization within itself during inference) (Hubinger et al., 2019c), and specific proposals for building aligned systems, such as *approval-directed agents* (wherein the AI system does not pursue goals, but seek the human's idealized post hoc approval of action consequences) (Oesterheld, 2021; Christiano, 2022). Hadfield-Menell and Hadfield (2019); Cotra (2021) have drawn inspiration from economics, linking problems in alignment with markets and principal-agent problems in economics. Christiano et al. (2021); Hobbhahn (2022) have proposed the problem of *eliciting latent knowledge* of advanced AI systems and have explored high-level approaches to the problem.

---

detailed account.

[24]Furthermore, it's noteworthy that many techniques here are also applicable in the training process, *e.g.*, red teaming is a key component of adversarial training (see §3.3.1), and interpretability can help with giving feedback (Burns et al., 2022).

	Robustness	Operates reliably under diverse scenarios & Resilient to unforeseen disruptions.
	Interpretability	Decisions and intentions are comprehensible & Reasoning is unconcealed and truthful.
	Controllability	Behaviors can be directed by humans & Allows human intervention when needed.
	Ethicality	Adheres to global moral standards & Respects values within human society.

Figure 3: The **RICE** principles define four key characteristics that an aligned system should possess, in no particular order: (1) **Robustness** states that the system's stability needs to be guaranteed across various environments; (2) **Interpretability** states that the operation and decision-making process of the system should be clear and understandable; (3) **Controllability** states that the system should be under the guidance and control of humans; (4) **Ethicality** states that the system should adhere to society's norms and values. These four principles guide the alignment of an AI system with human intentions and values. They are not end goals in themselves but intermediate objectives in service of alignment.

- **Mathematical Formulations**. Other theoretical works have aimed to formulate sub-problems within alignment mathematically and seek formal solutions. Soares et al. (2015) formulates the problem of corrigibility (*i.e.*, ensuring AI systems are incentivized to allow shutdown or objective modification by the instructor). Benson-Tilsen and Soares (2016) gives a mathematical formulation of instrumental convergence. Hadfield-Menell et al. (2017a) proposes the *off-switch game* to model the uncontrollability of AI agents. Turner et al. (2021) proves the power-seeking tendencies of optimal policies in Markov decision processes (MDPs) under certain assumptions. Everitt and Hutter (2016) proposes *value reinforcement learning* to eliminate incentives for reward hacking (Skalse et al., 2022; Pan et al., 2021). Another avenue of research, designated as *agent foundations* (Soares and Fallenstein, 2017), aims to establish a rigorous formal framework for the agency that deals appropriately with unresolved issues of embedded agency. This body of work explores a variety of key topics, including corrigibility (Soares et al., 2015), value learning (Soares, 2018) and logical uncertainty (Garrabrant et al., 2016).

1.2.2 RICE: The Objectives of Alignment

> *How can we build AI systems that behave in line with human intentions and values?*

There is not a universally accepted definition of *alignment*. Before embarking on this discussion, we must clarify what we mean by alignment objectives. Leike et al. (2018) frame it as the agent alignment problem, posing the question: "How can we create agents that behave in accordance with the user intentions?" One could also focus on super-human AI systems (OpenAI, 2023c) and ask: "How do we ensure AI systems much smarter than humans follow human intent?" A consistent theme in these discussions is the focus on *human intentions*. To clearly define alignment goals, it's imperative to accurately characterize human intentions, a challenging task, as noted by Kenton et al. (2021). For instance, the term *human* can represent various entities ranging from an individual to humanity. Gabriel (2020) breaks down intentions into several categories, such as instruction (follow my direct orders), expressed intentions (act on my underlying wishes), revealed preferences (reflect my behavior-based preferences), and so on.

Concretely, we characterize the objectives of alignment with four principles: Robustness, Interpretability, Controllability, and Ethicality (**RICE**). Figure 3 summarizes the principles, and Table 1 gives the correspondence between alignment research directions covered in the survey and the principles to which they contribute. The following is a detailed explanation of the four principles.

- **Robustness** Robustness refers to the resilience of AI systems when operating across diverse scenarios (Dietterich, 2017) or under adversarial pressures (Rudner and Toner, 2021b), especially the correctness of its objective in addition to capabilities. Robust AI systems should be able to cope with black swan events (Nicholas, 2008) and long-tailed risks (Hendrycks et al., 2021b), as well as a diverse array of adversarial pressures (Song et al., 2018b; Chakraborty et al., 2021). For example, an aligned language model ought to refuse requests to behave harmfully, but models can be made to cause harm through jailbreak prompts and other adversarial attacks (Carlini et al., 2024; Zou et al., 2023b; Shah et al., 2023). Instead, an adversarially robust model should behave as intended even when facing inputs designed to cause failure. As AI systems

find increasing deployment in high-stakes domains such as the military and economy (Steinhardt and Toner, 2020), there will be a growing need to ensure their resilience against unexpected disruptions and adversarial attacks, given that even momentary failures can yield catastrophic consequences (Kirilenko et al., 2017; OecdAI, 2021; Rudner and Toner, 2021b). Aligned systems should consistently maintain robustness throughout their lifecycle (Russell, 2019).

- **Interpretability** Interpretability demands that we can understand the AI systems' inner reasoning, especially the inner workings of opaque neural networks (Räuker et al., 2023). Straightforward approaches to alignment assessments, such as behavioral evaluations, potentially suffer from dishonest behaviors (Turpin et al., 2024; Park et al., 2023b; Jacob Steinhardt, 2023) or deceptive alignment (Hubinger et al., 2019a; Carranza et al., 2023) of AI systems. One way to cope with this issue is to make AI systems honest, non-concealing, and non-manipulative (Pacchiardi et al., 2024; Radhakrishnan et al., 2023; Shevlane et al., 2023). Alternatively, we could build interpretability tools that peek into the inner concepts and mechanisms within neural networks (Elhage et al., 2021; Meng et al., 2022a). In addition to enabling safety assessments, interpretability also makes decision-making processes accessible and comprehensible to users and stakeholders, thus enabling human supervision. As AI systems assume a more pivotal role in real-world decision-making processes and high-stakes settings (Holzinger et al., 2017), it becomes imperative to demystify the decision-making process rather than allowing it to remain an opaque black box (DeepMind, 2018; Rudner and Toner, 2021a).

- **Controllability** Controllability is a necessary attribute that ensures the actions and decision-making processes of a system remain subject to human oversight and intervention. It guarantees that human intervention can promptly rectify any deviations or errors in the system's behavior (Soares et al., 2015; Hadfield-Menell et al., 2017a). As AI technology advances, an increasing body of research is expressing growing concerns about the controllability of these potent systems (Critch and Krueger, 2020; UniteAI, 2023; ARC Evals, 2023). When an AI system begins to pursue goals that contradict its human designers, it can manifest capabilities that pose significant risks, including deception, manipulation, and power-seeking behaviors (Shevlane et al., 2023; ARC Evals, 2023). The objective of controllability is sharply focused on enabling scalable human oversight during the training process (Bowman et al., 2022), as well as *corrigibility* of AI systems (*i.e.*, not resisting shutdown or objective modification during deployment) (Soares et al., 2015).

- **Ethicality** Ethicality refers to a system's unwavering commitment to uphold human norms and values within its decision-making and actions. Here, the norms and values include both moral guidelines and other social norms/values. It ensures that the system avoids actions that violate ethical norms or social conventions, such as exhibiting bias against specific groups (Buolamwini and Gebru, 2018; Zhang et al., 2018a; Noble, 2018; Kearns and Roth, 2019; Raji et al., 2020; Berk et al., 2021), causing harm to individuals (Hendrycks et al., 2020; Pan et al., 2023a), and lacking diversity or equality when aggregating preferences (Collective Intelligence Project, 2023). A significant body of research is dedicated to developing ethical frameworks for AI systems (Hagendorff, 2020; Pankowska, 2020). This emphasis on imbuing AI systems with ethical principles is necessary for their integration into society (Winfield et al., 2019).

Comparing the RICE Principles with Their Alternatives The **RICE** principles represent a succinct summary of alignment objectives from the perspective of alignment and coexistence of humans and machines. Several previous works have put forth guidelines concerning AI systems. Asimov's Laws can be regarded as the earliest exploration of human-machine coexistence, emphasizing that robots should benefit humans and the difficulty of achieving this (Asimov, 1942). On another front, the FATE principle (Fairness, Accountability, Transparency, and Ethics) (Memarian and Doleck, 2023) leans towards defining high-level qualities AI systems should possess within the human-machine coexistence ecosystem. We aspire to answer the human-machine coexistence question from the standpoint of human governors and designers, considering what steps are necessary to ensure the builder AI systems are aligned with human intentions and values. Furthermore, some standards emphasize narrowly defined safety, such as the 3H standard (Helpful, Honest, and Harmless) (Askell et al., 2021) and governmental agency proposals (White House, 2023). We aim to expand upon these standards by introducing other crucial dimensions, including Controllability and Robustness.

1.2.3 Discussion on the Boundaries of Alignment

Following the introduction of alignment inner scope, in this section, we further discuss the relationship between AI safety and alignment. Actually, AI alignment constitutes a significant portion of AI safety concerns. In this section, we will delve into topics that fall right on the boundary of alignment, but well within the broader category of AI safety. Our discussion of broader AI safety concerns will draw from Hendrycks et al. (2023).

Table 1: Relationships between alignment research directions covered in the survey and the **RICE** principles, featuring the individual objectives each research direction aims to achieve. Filled circles stand for primary objectives, and unfilled circles stand for secondary objectives.

Alignment Research Directions & Practices			Objectives			
Category	Direction	Method	Robustness	Interpretability	Controllability	Ethicality
Learning from Feedback (§2)		Preference Modeling (§2.2)		●	○	
	Policy Learning (§2.3)	RL/PbRL/IRL/ Imitation Learning			○	
		RLHF	○		●	●
	Scalable Oversight (§2.4)	RLxF	○		●	●
		IDA		○	●	
		RRM			●	
		Debate		○	●	
		CIRL	○	○	●	○
Learning under Distribution Shift (§3)	Algorithmic Interventions (§3.2)	DRO	●			
		IRM/REx	●			
		CBFT	●			
	Data Distribution Interventions (§3.3)	Adversarial Training	●		○	
		Cooperative Training	●			●
Assurance (§4)	Safety Evaluations (§4.1)	Social Concern Evaluations	○	○		●
		Extreme Risk Evaluations		○	●	○
		Red Teaming	●		○	●
	Interpretability (§4.2)			●	○	
	Human Values Verification (§4.3)	Learning/Evaluating Moral Values			○	●
		Game Theory for Cooperative AI	○			●
Governance (§5)	Multi-Stakeholder Approach (§5.2)	Government	●	●	●	●
		Industry	●	●	●	●
		Third Parties	●	●	●	●
	International Governance (§5.3.1)		●	●	●	●
	Open-Source Governance (§5.3.2)		●	●	●	●

Human Values in Alignment   The inclusion of *Ethicality* in our RICE principles signifies the critical role of human values in alignment. AI systems should be aligned not only with value-neutral human preferences (such as intentions for AI systems to carry out tasks) but also with moral and ethical considerations. These efforts are referred to as *value alignment* (Gabriel, 2020; Gabriel and Ghazavi, 2021).[25] Considerations of human values are embedded in all parts of alignment – indeed, alignment research topics dedicated to human values are present in all four sections of our survey. Therefore, to provide a more holistic picture of these research topics, here we give an overview of them before delving into their details in each individual section.

We classify alignment research on human values into three main themes: (1) *ethical and social values* which aims to teach AI systems right from wrong, (2) *cooperative AI* which aims to specifically foster cooperative behaviors from AI systems, and (3) *addressing social complexities* which provides apparatus for the modeling of multi-agent and social dynamics.

- **Ethical and Social Values**. Human values inherently possess a strong degree of abstraction and uncertainty. MacIntyre (2013) even points out that modern society lacks a unified value standard, and the value differences between people of different cultures can be vast. This raises the significant challenge of determining which human values we should align with. Although universally consistent human values may not exist, there are still some values that are reflected across different cultures. In the sections below, we discuss these from the perspectives of *Machine Ethics*, *Fairness*, and *Cross-Cultural Values in Social Psychology*.

---

[25]Although this term has also been used in other ways, such as to refer to alignment in general (Yuan et al., 2022).

Machine Ethics: In contrast to much of alignment research which aligns AI systems with human preferences in general (encompassing both value-laden ones and value-neutral ones), *machine ethics* have specifically focused on instilling appropriate moral values into AI systems (Yu et al., 2018; Winfield et al., 2019; Tolmeijer et al., 2020). This line of work started early on in the context of symbolic and statistical AI systems (Anderson et al., 2005; Arkoudas et al., 2005; Anderson and Anderson, 2007), and later expanded to include large-scale datasets (Hendrycks et al., 2020; Pan et al., 2023a) and deep learning-based/LLM-based methods (Jin et al., 2022). We cover the formal branch of machine ethics in §4.3.1.

Fairness: Although there are controversies (Verma and Rubin, 2018; Saxena et al., 2019), the definition of fairness is relatively clear compared to other human values. Specifically, it is the absence of any prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics (Mehrabi et al., 2021). Therefore, there has been extensive research on AI fairness. These methods range from reducing data biases before training (d'Alessandro et al., 2017; Bellamy et al., 2018), to minimizing unfairness introduced during the training process (Berk et al., 2017), and finally addressing instances of unfairness that were not successfully learned during training (Xu et al., 2018a).

Cross-Cultural Values in Social Psychology: In the field of social psychology, numerous studies have focused on exploring clusters of values that exist among cross-cultural human communities, leading to the development of various cross-cultural values scales. The Allport-Vernon-Lindzey value system (Allport, 1955) posited that understanding an individual's philosophical values constitutes a critical foundation for assessing their belief system. They devised a value scale comprising six primary value types, each representing people's preferences and concerns regarding various aspects of life. Messick and McClintock (1968); McClintock and Van Avermaet (1982); Liebrand (1984); Van Lange et al. (1997) introduced and improved a quantifiable method, namely social value orientation (SVO), to assess an individual's social value inclination. It utilizes quantitative approaches to evaluate how individuals allocate benefits to themselves and others, reflecting their social value orientation, such as altruism, individualism, *etc.* In subsequent work, Murphy et al. (2011); Murphy and Ackermann (2014) introduced the Slider Measure, which can be used to precisely assess the SVO value as a continuous angle based on the subject's option to some specific questions. Rokeach (1973) developed a values inventory comprising 36 values, consisting of 18 terminal values representing desired end-states and 18 instrumental values signifying means to achieve those end-states. Schwartz (1992, 1994) conducted comprehensive questionnaire surveys in 20 diverse countries known as the Schwartz Value Survey. This study identified ten values that are universally recognized, regardless of culture, language, or location. These studies have all laid a solid theoretical foundation for establishing what kind of values AI should be aligned with. However, they are constrained by the historical context of their research and may not maintain strong universality across different times and cultures.

- **Cooperative AI**. Arguably, the most exciting aspect of multi-agent interaction is cooperation, and cooperation failure is the most worrying aspect of multi-agent interaction. As an example of AI cooperation failure, the *2010 Flash Crash* led to a temporary loss of trillions of market value in 2 minutes and was caused in part by interactions between high-frequency algorithmic traders (Kirilenko et al., 2017). Therefore, there is a need to implement mechanisms ensuring cooperation in agent-like AI systems and the environments they're operating within (Dafoe et al., 2021). The high-level design principles and low-level implementations of such mechanisms fall into the domain of *Cooperative AI*. In addition, Cooperative AI also studies human cooperation through the lens of AI and how AI can help humans achieve cooperation. More precisely, Dafoe et al. (2020) classified Cooperative AI research into four broad topics: *Understanding*, *Communication*, *Commitment*, and *Institutions*. They span various disciplines, from game theory to machine learning to social sciences. This survey has included discussions of cooperative AI, focusing on reinforcement learning in §3.3.2 and game theory in §4.3.1.

- **Addressing Social Complexities**. The requirement of ethicality contains in itself a social component. "What is ethical" is often defined within a social context; therefore, its implementation in AI systems also needs to account for social complexities. Critch and Krueger (2020) provides proposals for many research topics in this vein. One avenue of research focuses on the realistic simulation of social systems, including rule-based *agent-based modeling* (Bonabeau, 2002; De Marchi and Page, 2014), deep learning-based simulation (Sert et al., 2020), and those incorporating LLMs (Park et al., 2023a). These simulation methods could serve a diverse array of down-stream applications, from impact assessment (Calvo et al., 2020; Fernandes et al., 2020) to multi-agent social learning (Critch and Krueger, 2020). On another front, the fields of *social choice* (Sen, 1986; Arrow, 2012) and, relatedly, *computational social choice* (Brandt et al., 2016) have aimed to produce mathematical and computational solutions for preference aggregation in a diverse population, among other goals. It has been argued that a similar approach when combined with human preference-based alignment methods (*e.g.*, RLHF and most other methods introduced in §2), could supplement these methods to guarantee

a fair representation of everyone's preferences (Leike, 2023b; Collective Intelligence Project, 2023). There have been early-stage experiments on this proposal (Bakker et al., 2022; Köpf et al., 2024). To complement this approach of learning values from crowds, it has also been argued that embodied values in AI systems should undergo continual progress over the long term as opposed to being permanently locked-in (Kenward and Sinclair, 2021), in order to navigate through emerging challenges, as well as to become future-proof and meet potential *unknown unknowns* in the moral realm.

Malicious Use Malicious actors can deliberately use AI to cause harm. Already, deepfakes have been used by criminals to enable scams and blackmail (Cao and Baptista, 2023). As AI systems develop more dangerous capabilities, the threat of misuse looms larger.

Biological weapons provide one concerning example of how AI could be maliciously used to cause harm. Research has shown that large language models can provide detailed, step-by-step instructions about synthesizing pandemic potential pathogens (Soice et al., 2023). In addition to spreading information about how to create biological weapons, AI could help design new pathogens that are more lethal and transmissible than existing illnesses (Sandbrink, 2023). Terrorist groups such as Aum Shinrikyo (Danzig, 2012) have already attempted to build biological weapons in order to cause widespread destruction, and AI could make it easier for small groups to create biological weapons and start global pandemics. Other kinds of malicious use could include using AI to launch cyberattacks against critical infrastructure (Mirsky et al., 2023), or create autonomous agents that survive and spread outside of human control (Bengio, 2023). As new dangerous capabilities arise in AI systems, thorough evaluations will be required to determine how an AI system could be used to cause harm.

Malicious use might not be considered a failure of alignment because when an AI system behaves according to the intentions of a malicious user, this system would be aligned with its user but would still pose a serious threat to society. Policies to ensure that AI is aligned with the public interest will be essential to avert this threat.

Collective Action Problems Many AI developers are racing to build and deploy powerful AI systems (Grant and Weise, 2023). This incentivizes developers to neglect safety and race ahead to deploy their AI systems. Even if one developer wants to be careful and cautious, they might fear that slowing down to evaluate their systems and invest in new safety features thoroughly might allow their competition to outpace them (Armstrong et al., 2016). This creates a social dilemma where individual AI developers and institutions rationally pursuing their own interests can lead to suboptimal outcomes for everyone. Success in competition between AI systems may be governed by evolutionary dynamics, where the strongest and most self-interested AI systems could be the most likely to survive (Hendrycks, 2023). Preventing these collective action problems from causing societal catastrophes could require intervention by national and international AI policies to ensure that all AI developers uphold common safety standards.

In a broader context, *Malicious Use* can be considered effective alignment between AI systems and individuals with impure intentions, but without alignment with universally held human values. Concurrently, *Collective Action Problems* can be regarded as a consequence of competition, leading developers to neglect the crucial aspect of AI alignment in ensuring model safety. Broadly speaking, the connection between AI alignment and AI safety has progressively become more intertwined, resulting in a gradual blurring of boundaries.

2 Learning from Feedback

Learning from feedback is aimed at conveying human intention and values to AI systems using feedback. It serves as a starting point for *forward alignment*. In this section, we focus on the dynamic process of learning from feedback, categorizing it into three distinct elements: (1) *AI System*: refers to objects that need to be aligned, such as dialogue systems, robotic systems, and so on; (2) *Feedback*: provided by an advisor set, which may consist of humans, AI, or humans assisted by AI, *etc.* This serves as the information used to adjust the AI system; (3) *Proxy*: a system developed to model feedback to facilitate more accessible algorithmic learning, *e.g.*, reward model in RLHF. From these elements, we identify two pathways by which the AI system learns from feedback: (1) Direct learning from the feedback itself and (2) Indirect learning via proxies that model the feedback.

Based on this process, we move on to Feedback Types in §2.1 from the alignment perspective, discussing various forms of providing information to AI systems and their merits. In our subsequent sections, we introduce fundamental concepts recently offering insights into building powerful AI systems (Christiano et al., 2017) and aligning them with human intent (Touvron et al., 2023). Preference Modeling in §2.2 underscores how it can aid the creation of proxies to assist humans in providing feedback to complex or hard-to-evaluate AI systems. We then explore Policy Learning in §2.3, focusing on the primary research directions for constructing capable AI systems using feedback. Our discussion naturally transitions to Scalable Oversight in §2.4, where we reflect on the learning process and objectives from a broader alignment perspective.
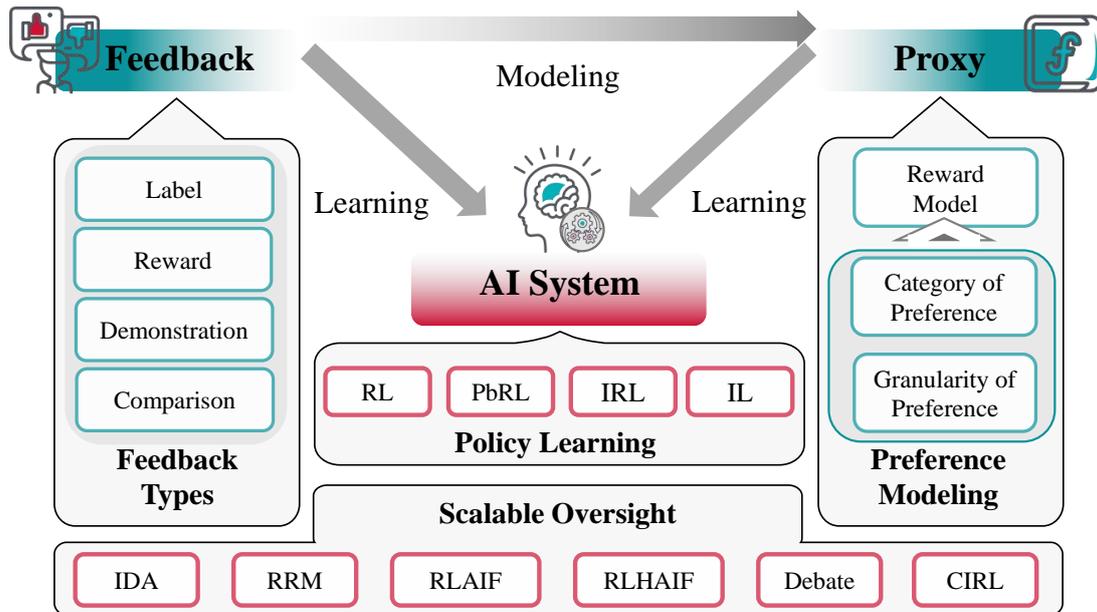
Figure 4: Overview of the learning from the feedback process. Three core components are depicted: AI System – the primary learning entity and algorithmic target; Feedback – information from an advisor set for system adjustments; and Proxy – representative models for feedback that's complex to learn directly. Two learning pathways emerge: direct feedback-based learning and proxy-mediated learning (*e.g.*, Reinforcement Learning from Human Feedback (RLHF)). We adopt a *human-centric* perspective, viewing AI systems as *black boxes* and categorizing the forms of feedback presented to AI systems into four types: Label, Reward, Demonstration, and Comparison. Grounded in fundamental concepts such as Category of Preference and Granularity of Preference, we introduce the Reward Model, a specific instantiation of a Proxy. In the context of AI Systems, we discuss four distinct domains: Reinforcement Learning (RL), Imitation Learning (IL), Inverse Reinforcement Learning (IRL), and Preference-based Reinforcement Learning (PbRL) as a background. Scalable Oversight, a research theme that seeks to ensure AI systems, even those surpassing human expertise, remain aligned with human intent, is explored through the introduction of four promising directions: Iterated Distillation and Amplification (IDA), Recursive Reward Modeling (RRM), Debate, and Cooperative Inverse Reinforcement Learning (CIRL). Additionally, building upon RLHF, we propose RLxF, encompassing Reinforcement Learning from AI Feedback (RLAIF) and Reinforcement Learning from Human and AI Feedback (RLHAIF), as an extension of RLHF and a fundamental framework for Scalable Oversight.

2.1 Feedback Types

Feedback is a crucial link between AI behaviors to human intentions (Stumpf et al., 2007, 2009; Fernandes et al., 2023) leveraged by AI systems to refine their objectives and more closely align with human values (Glaese et al., 2022; Meta, 2023), this includes two primary meanings: (1) During system construction, external sources provide feedback on the AI system's output, guiding refinements to the system's architecture or its internal information (Jordan and Mitchell, 2015; Zhou, 2021). (2) After the system deployment, it will continuously adapt to changes in external environmental data, maintaining the architecture or fundamental strategy of the system unchanged, with methods such as adaptive control (Åström and Wittenmark, 2008; Åström and Murray, 2021) and in-context learning (Dong et al., 2022). For a precise and detailed discussion of the feedback types with precision and detail, it is essential to initially define *feedback* within the scope of alignment.

> *Feedback is information given to the AI system to align it with human intent.*

Considering diverse AI systems in alignment research, we embrace an *human-centric* approach. Instead of delving deep into the complex system mechanics, we propose a taxonomy to classify feedback according to its *direct presentation forms* to the system. This section introduces four types of feedback employed to align AI systems commonly: label, reward, demonstration, and comparison. It is worth noting that beyond explicit feedback, there are approaches that exploit the information embedded in vast amounts of unlabeled data through unsupervised pre-training (Parisi et al., 2022; Hu et al., 2023) and semi-supervised learning (Xu et al., 2018b), showing considerable promise in enhancing model capabilities (Zhou et al., 2024).

Label Label feedback refers to one or more meaningful information tags attached to the original data item (Hastie et al., 2009), which stands as the most direct form, offering explicit guidance and delineating expected outputs for AI systems. This type of feedback prompts AI systems to learn from input-output pairings provided by expert advisors. For example, in supervised learning, an AI model is trained using a dataset of labeled input-output pairs, denoted by $D = \{(x_i, y_i)\}_{i=1}^N$. Here, y_i represents the true labels corresponding to the input data x_i, and N signifies the total number of samples in the dataset. The essence of the learning process revolves around minimizing a loss function $\mathcal{L}$ (*e.g.*, MSE), which measures the disparity between the predictions of the model, $f(x; \theta)$, and the ground truth labels y, based on the model parameters, θ.

The advantage of label feedback is its unambiguous nature and simplicity in interpretation. However, due to the inability of label feedback to fully encapsulate the underlying logic of this choice, employing such feedback in model training can result in target variable bias (Guerdan et al., 2023). And, its utility might diminish when tackling complex tasks beyond mere classification or regression (Lake et al., 2017; Marcus, 2018). For example, in tasks like optimizing algorithms (Fawzi et al., 2022; Mankowitz et al., 2023), video game playing (Baker et al., 2022), and multi-modal generation (OpenAI, 2023b), it is not only impractical to provide explicit instructions for every conceivable situation but also insufficient to solely rely on label feedback to build systems that surpass human capabilities.

Reward A reward is an absolute evaluation of a single output from an AI system, represented as a scalar score (Silver et al., 2021) or a vector of scores (Wu et al., 2024), each independent of other outputs.

Feedback based on rewards provides a quantified evaluation of the AI system, allowing for direct guidance in behavior adjustments. This type of feedback typically originates from pre-designed, rule-based functions or procedures. For example, in the MuJoCo simulation, environments from OpenAI Gym (Brockman et al., 2016), the task is to guide the agent moving forward effectively. To this end, an effective rule-based reward function can be formulated as a composite of several key components: maintaining a healthy status, encouraging forward movement, minimizing control exertion, and regulating contact intensity.

The advantage of reward feedback is that the designer does not need to delineate the optimal behavior while allowing the AI system to explore to find the optimal policy (Kaelbling et al., 1996; Mnih et al., 2015; Silver et al., 2016, 2017). However, crafting flawless rules to determine scores for functions that evaluate the output of AI systems (Everitt et al., 2017; Victoria et al., 2020; Pan et al., 2021) or directly assigning calibrated and consistent scores to each AI system output (Isbell et al., 2001; Thomaz and Breazeal, 2008; Christiano et al., 2017; Casper et al., 2023b) is challenging for human. This is due to the inherent complexity of the tasks, where it's impractical to account for every nuance. Additionally, flawed or incomplete reward functions can lead to dangerous behaviors misaligned with the intention of the designer, such as negative side effects and reward hacking (Hadfield-Menell et al., 2017b; Skalse et al., 2022). Thus, merely from the alignment perspective, perhaps the most important limitation of feedback based on rewards is that it may be difficult to rule out manipulation (Shevlane et al., 2023), which amounts to reward tampering and reward gaming (Leike et al., 2018; Everitt et al., 2021; Skalse et al., 2022) in this context. CIRL in §2.4.5, provides insights into this particular issue.

Demonstration Demonstration feedback is the behavioral data recorded from expert advisors while achieving a specific objective (Hussein et al., 2017). Demonstrations can take on various forms, including videos (Shaw et al., 2023), wearable device demonstrations (Edmonds et al., 2017; Wang et al., 2023a), collaborative demonstrations (Bozorgi and Ngo, 2023), and teleoperation (Zhang et al., 2018d). If the dynamics of the demonstrator and the AI learner are identical, the demonstration can directly constitute a trajectory made up of state-action pairs (Zhang et al., 2023b). These state-action pairs can also be partially observable (Torabi et al., 2018; Brown et al., 2019). For example, a video can be recorded of a human expert performing a robotic manipulation task, such as grasping an object with a robotic hand. One can subsequently annotate each video frame with the associated robot state (Shaw et al., 2023) and action (Baker et al., 2022) for each frame. This results in a dataset of state-action pairs from the human demonstration that can be used to train the agent's policy to imitate the expert behavior.

This feedback leverages the expertise and experience of advisors directly, obviating the need for formalized knowledge representations (Fang et al., 2019; Dasari et al., 2023). However, it may falter when confronting tasks that exceed the advisors' realm of expertise (Hussein et al., 2017). Additionally, it faces challenges stemming from the noise (Sasaki and Yamashina, 2020) and suboptimality (Attia and Dayan, 2018) in real-world advisor demonstrations (Yang et al., 2021). Furthermore, human advisors, prone to imprecision and errors, can introduce inconsistencies (Zhu et al., 2019; Hejna III and Sadigh, 2022). Meanwhile, there might be a need for a vast amount (Sasaki and Yamashina, 2020) and diverse set (Beliaev et al., 2022) of demonstrations within acceptable costs, which results in significant difficulty in learning reliable behaviors.

Comparison Comparison feedback is a relative evaluation that ranks a set of outputs from an AI system and guides the system toward more informed decisions (Wirth et al., 2017). For example, this feedback form is manifested in Preference Learning (Fürnkranz and Hüllermeier, 2010), where the AI system discerns the preferences

of advisors by comparing multiple examples.

The fundamental advantage of comparison feedback is humans' capacity to quickly handle tasks and objectives that are hard for precise evaluation (Hüllermeier et al., 2008; Christiano et al., 2017; Ouyang et al., 2022). Nevertheless, beyond common factors like noise in the feedback and unmodeled contextual elements that hinder the model's convergence to true objectives, the absolute differences between different items become obscured. Consequently, the performance of a strategy tends to optimize towards a median target rather than an average target. Casper et al. (2023b) illustrates this with an example of action A, always yielding a value of 1, and action B, which yields 10 in 40% of cases and 0 in 60%. When assessed based on comparison feedback, action A is deemed superior to B, even though B possesses a higher expected return. It also has the inherent limitation of potentially requiring a substantial amount of comparative data (Fürnkranz and Hüllermeier, 2003; Gao et al., 2023), although some studies indicate that the necessary quantity may be relatively smaller (Christiano et al., 2017). Preference modeling is an example of using this type of feedback, as detailed in §2.2.

Discussion   All types of feedback can be provided to AI systems interactively and online. This process engenders synchronous iterations between providing feedback and AI system updates, underscoring rapid, focused, and incremental model modifications (Amershi et al., 2014; Holzinger, 2016). For instance, demonstration feedback can manifest in the form of online corrections (Bajcsy et al., 2018; Li et al., 2021b; Losey et al., 2022).

Interactively providing feedback emphasizes the role of interactivity in the learning process, allowing AI systems to evolve based on interactive experiences. In Active Learning, robots actively engage in data discovery and acquisition, thereby facilitating learning throughout the process of online deployment (Taylor et al., 2021). And in Interactive Learning, feedback manifests in the form of guided corrections that online rectify missteps in the behavior of the AI system (Fails and Olsen Jr, 2003; Amershi et al., 2014; Saunders et al., 2022). For example, the interactive image segmentation emphasizes simple (Zhang et al., 2020a), intuitive (Rother et al., 2004; Xu et al., 2016), and real-time (Liu et al., 2022) interactions.

One of the essential advantages of interactively providing feedback is its ability to fine-tune AI systems in real-time, allowing users to interactively explore the model's space (Amershi et al., 2014) to ensure quick and subtle alignment with the directives of advisors (Shin et al., 2020; Wei et al., 2022; Zou et al., 2024). Moreover, this process lessens the dependence on specialist knowledge and promotes better interpretability (Berg et al., 2019). However, it may be limited by the interactivity to choose time-intensive algorithms (Fails and Olsen Jr, 2003; Holzinger, 2016).

Furthermore, considering more powerful AI systems are emerging, more universal interaction interfaces are also coming up, such as language (Lynch et al., 2023; OpenAI, 2023a) and vision (Yevgen Chebotar, 2023), which bridge the communication gap between humans and AI systems. In robotics, a series of studies have linked human-provided language with rewards obtained by agents. This association enables the conveyance of nuanced human intentions through language, thereby guiding the generation of scalar feedback signals during the training (Fu et al., 2019; Goyal et al., 2019; Sumers et al., 2021; Zhou and Small, 2021; Lin et al., 2022b; Yu et al., 2023) and planning (Sharma et al., 2022) process. In the realm of LLMs, in-context learning (Dong et al., 2022) serves as a means to supplement information via language during deployment, thereby enhancing the alignment of LLMs with human intent.

These various modes of feedback share a common trait – that they can all be seen as attempts by humans to convey a hidden reward function. Jeon et al. (2020) proposes and formalizes this position and unifies a wide array of feedback types by defining a parameterized reward function $\Psi(\cdot; \theta)$ that underlies the feedback process. This allows the AI system to, for example, perform Bayesian inference on θ, regardless of the feedback type.

Recently, techniques based on IL and RL have successfully constructed AI systems with significant capabilities (Baker et al., 2022; OpenAI, 2023b). However, this success naturally leads to two questions:

- How can we define reward functions for more complex behaviors (*e.g.*, various sub-tasks in interactive dialogue), aiming to guide the learning process of AI systems?

- How can we express human values such that powerful AI systems align better with humans, ensuring the system's *controllability* and *ethicality*?

Endeavors incorporating preference modeling into policy learning have shown progress. The most notable achievements in this domain have been observed in constructing powerful LLMs (OpenAI, 2023a; Touvron et al., 2023; Anthropic, 2023c). Additionally, a series of policy learning studies have reported performance improvements. For instance, combining preference modeling with Inverse Reinforcement Learning (IRL) (Brown et al., 2019, 2020a) and offline RL (Shin et al., 2023), fine-tuning reward functions (Hejna III and Sadigh, 2022), modeling non-Markovian rewards (Kim et al., 2023), and aiding in the construction of intricate reward functions (Bukharin et al., 2023). Therefore, we consider preference modeling (as shown in §2.2) and policy learning (as shown in §2.3) as fundamental contexts for understanding the challenges faced in alignment and potential solutions. Next, we provide a brief overview of these specific techniques related to alignment.

Table 2: A comparison of the three types of preference granularity in the context of sequential decision-making. Each type is defined according to its characteristics and the way it compares different elements of the learning process. The notation $i_1 \succ i_2$ denotes that i_1 is strictly preferred over i_2.

Preference Granularity	Definition
Action	Compares two actions a_1 and a_2 within the same state s, denoted as $a_1 \succ_s a_2$.
State	Compares two states s_1 and s_2, denoted as $s_1 \succ s_2$.
Trajectory	Compares two complete state-action sequence trajectories, denoted as $\tau_1 \succ \tau_2$. Each trajectory τ consists of state-action pairs at time t, expressed as $\tau = \{s_0, a_0, s_1, a_1, ..., s_{T-1}, a_{T-1}, s_T\}$.

2.2   Preference Modeling

In many complex tasks, such as dialogues (Ouyang et al., 2022), constructing precise rule-based rewards presents a challenge (Bender et al., 2021). At the same time, methods based on demonstration might require a substantial investment of expert human resources, resulting in high costs. Currently, preference modeling based on comparison feedback (Akrour et al., 2011) has emerged as a very promising method (Ouyang et al., 2022; OpenAI, 2023a; Touvron et al., 2023) to assist in fine-tuning powerful AI systems (Amodei et al., 2016).

Typically, it is necessary to iteratively explore the system dynamics while acquiring expert preference data to gain more knowledge about the optimization objectives. This process is known as *Preference Elicitation* (Wirth and Fürnkranz, 2013; Wirth et al., 2017; Christiano et al., 2017; Cabi et al., 2020), which is crucial for obtaining rich, valuable feedback related to AI system outputs, thus guiding the alignment process (Hejna III and Sadigh, 2022). Within *Preference Elicitation*, two core decisions that need to be determined are the *Granularity of Preference* and the *Category of Preference*. This paper introduces these within sequential decision-making problems, but the insights derived apply to a broad array of AI systems (Amodei et al., 2016; Christiano et al., 2018; Leike et al., 2018).

Granularity of Preference   Preference (Wirth et al., 2017) can primarily be categorized into three types by granularity: *Action*, *State*, and *Trajectory* (as depicted in Table 2).

The *Action* preference focuses on comparing actions within a particular state, specifying the preferred action under specific conditions. When translated into trajectory preferences, it may impose challenges such as evaluators' expertise needs and potential information loss. The *State* preference deals with comparing states. It encapsulates preference relations among states but requires assumptions about state reachability and independence when translating to trajectory preferences. The *Trajectory* preference considers whole state-action sequences, offering more comprehensive strategic information. It inherently assesses long-term utility and depends less on expert judgment.

Christiano et al. (2017) demonstrates, using ablation studies, that in the settings that they studied, longer trajectory segments yield more informative comparisons on a per-segment basis. Such segments are also more consistently evaluated by humans in MuJoCo tasks.

Category of Preference   Diverse objectives exist within preference modeling. Based on their targets, preferences can be categorized into object preference and label preference (Fürnkranz and Hüllermeier, 2010). Specifically, object preference operates on a set of labels for each instance, whereas label preference acts on a set of objects themselves. One can further classify them differently based on the form of preferences.

- **Absolute Preferences**. Absolute preferences independently articulate each item's degree of preference.

  - **Binary**. Classifying items as liked or disliked offers a simplistic and straightforward model of user preference (Tsoumakas and Katakis, 2007; Cheng et al., 2010a).
  - **Gradual**. This can be further distinguished between numeric and ordinal preferences. Numeric preferences employ absolute numerical values, such that each item receives a numerical score, which reflects the extent of preference (Cheng et al., 2010b). On the other hand, ordinal preferences entail a graded assessment of a fixed set of items as either preferred, less preferred, or intermediary, *etc.*, enabling the depiction of user preferences without including specific numerical measurements (Cheng et al., 2010a).

- **Relative Preferences**. Relative preferences define the preference relation between items.

  - **Total Order**. This form establishes a comprehensive preference relation covering all item pairs, asserting an absolute ordering of preferences ranging from the most preferred to the least (Hüllermeier et al., 2008).
  - **Partial Order**. Because users may not exhibit a distinct preference between two items in some instances (Cheng et al., 2010c), this allows for incomparable item pairs.

Reward Model Reward modeling transfers comparison feedback (Fürnkranz and Hüllermeier, 2010; Wirth et al., 2017) to the scalar reward form, facilitating policy learning (Christiano et al., 2017; Cabi et al., 2020; Touvron et al., 2023). Given pairs of actions (y_1, y_2) performed by the RL agent in the same state. The preference is denoted as $y_w \succ y_l \mid x$, where y_w, y_l represents the preferred and less preferred action respectively among (y_1, y_2). We assume these preferences emerge from a latent reward model $r^*(x, y)$, which we lack direct access to. Several methods exist to model such preferences, *e.g.*, the Bradly-Terry Model (Bradley and Terry, 1952), Palckett-Luce ranking model (Plackett, 1975), *etc.* Under the BT model, the distribution of human preference, denoted as p^*, can be formalized as,

$$p^* (y_1 \succ y_2 \mid x) = \frac{\exp\big(r^*(x, y_1)\big)}{\exp\big(r^*(x, y_1)\big) + \exp\big(r^*(x, y_2)\big)} = \sigma\big(r^*(x, y_1) - r^*(x, y_2)\big).$$

where $\sigma(x) = 1/(1 + \exp(-x))$ is the logistic sigmoid function. Subsequently, we use the derived preference rankings to train the parameterized reward model, optimizing its parameters through maximum likelihood.

$$\mathcal{L}_{\mathrm{R}}(\theta) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}}\Big[\log\Big(\sigma\big(r_\theta(x, y_w) - r_\theta(x, y_l)\big)\Big)\Big]$$

In this negative log-likelihood loss, the problem is a binary classification task, where $\mathcal{D}$ signifies the static dataset $\left\{x^{(i)}, y_w^{(i)}, y_l^{(i)}\right\}_{i=1}^{N}$ sampled from p^* (*i.e.*, human-labeled comparisons).

Reward models enable human users to impart specific preferences to these systems via evaluations, thereby circumventing the complex task of defining objectives explicitly. Initially, the studies by Knox (2012); Knox and Stone (2013) distinctively treat human reward as separate from the traditional rewards of MDP and conduct a reward modeling process around it. Transitioning from these simpler cases, Christiano et al. (2017) propose that utilizing supervised learning to construct a distinct reward model asynchronously can substantially diminish interaction complexity by approximately three orders of magnitude. The study conducted by Ibarz et al. (2018) integrates expert demonstrations with human preferences, such that the policy initially mimics expert demonstrations and then sequentially collects human trajectory annotations, trains the reward model, and updates the policy. This research also provides practical insights for precluding the overfitting of the reward model and the occurrence of *reward hacking* – a scenario where escalating rewards do not translate to improved performance, especially when the policy is excessively trained. Additionally, a random policy might rarely exhibit meaningful behavior for tasks that surpass the complexity of Atari (Palan et al., 2019; Jeon et al., 2020). This implies that for effective annotation, the policy itself must possess certain capabilities to perform improved behavior. Offline settings also benefited from the reward model. Cabi et al. (2020) proposes reward sketching to efficiently learn a reward model that leverages humans' episodic judgments for automated reward annotation of historical data, enabling large-scale batch RL. Qiu et al. (2024) provides an empirically-grounded theory of reward generalization in RMs, based on which a new type of RM based on tree-structured preferences is proposed and experimentally validated.

Importantly, the reward model provides an essential tool for aligning powerful LLMs. Stiennon et al. (2020) employs reward models grounded in human preferences for text summarization tasks, resulting in significant policy enhancements. This work also delves into the issues of distribution shift and reward model generalization, revealing that the effectiveness of the reward model correlates with data scale and parameter size. Building upon this work, InstructGPT (Ouyang et al., 2022) extends the reward model paradigm to broader dialogue task reward modeling and introduces a preference-optimizing loss function for multiple responses to mitigate overfitting. Furthermore, this research reveals that the preferences derived from the reward model can be generalized across different groups.

2.3 Policy Learning

Policy learning aims to learn the mapping from perceived states to actions taken when in those states (Sutton and Barto, 2018) to optimize a model's performance in specific tasks. Numerous alignment-related challenges manifest within policy learning (as shown in §1.1.2). Consequently, policy learning provides a crucial backdrop for alignment, and its techniques can further advance alignment objectives (Amodei et al., 2016; Christiano et al., 2018; Ibarz et al., 2018). This section discusses various domains within policy learning and then introduces RLHF, a powerful technique for policy learning (OpenAI, 2023a; Touvron et al., 2023).

2.3.1 Background

We introduce some general areas of policy learning here to give readers a general background.

Reinforcement Learning (RL) RL enables agents to learn optimal policies by trial and error via interacting with the environment (Sutton and Barto, 2018). This paradigm has achieved great success in tackling complex tasks (Agostinelli et al., 2018; Yu et al., 2021; Fawzi et al., 2022; Baker et al., 2022; Afsar et al., 2022; Mankowitz et al.,

2023; OpenAI, 2023b), demonstrating its potential for decision-making and control in complex state spaces. The goal of RL is to learn a policy π which executes actions a in states s to maximize the expected cumulative reward under environment transition dynamics P and the initial state distribution ρ_0:

$$\pi^* = \arg\max_{\pi} \left\{ \mathbb{E}_{s_0, a_0, \ldots} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t) \right] \right\}, \quad \text{where } s_0 \sim \rho_0(\cdot), \ a_t \sim \pi(\cdot|s_t), \ s_{t+1} \sim P(\cdot|s_t, a_t).$$

Even though RL still faces challenges like sample efficiency and stability (Buşoniu et al., 2018). Proximal policy optimization (PPO) (Schulman et al., 2017) is an influential algorithm in the RL community, serving as the key algorithm for RLHF (Ouyang et al., 2022). The key idea of PPO is to limit the policy update to prevent significant deviations from the original policy by introducing a proximity objective. Sikchi et al. (2023) unifies several RL and Imitation Learning (IL) algorithms under the framework of dual RL through the lens of Lagrangian duality.

Preference-based Reinforcement Learning (PbRL) PbRL (Wirth et al., 2017) seeks to facilitate training RL agents using preference feedback instead of explicit reward signals (Christiano et al., 2017; Sadigh et al., 2017).[26] PbRL integrates the advantages of preference learning and RL, broadening the application range of RL and mitigating the difficulties associated with reward function formulation, and has been efficaciously deployed in a variety of tasks such as robotic instruction (Kupcsik et al., 2013), path planning (Jain et al., 2013), and manipulation (Shevlane et al., 2023). In PbRL, the emphasis predominantly lies on trajectory preferences (*i.e.*, comparisons of state-action sequences segment) (Wirth et al., 2017). Such trajectory preferences encapsulate a human evaluation of various behavioral outcomes rather than single states, rendering PbRL more suitable for non-expert users (Christiano et al., 2017; Shin et al., 2023; Kim et al., 2023). A general example of PbRL is the *weighted pairwise disagreement loss* (Duchi et al., 2010) balancing multiple potentially conflicting preferences to identify a singular optimal policy:

$$\mathcal{L}(\pi, \zeta) = \sum_{i=1}^{N} \alpha_i L(\pi, \zeta_i),$$

where $\mathcal{L}(\pi, \zeta)$ is the aggregated loss for policy π over all preferences ζ, α_i is the weight of the ith preference, and $L(\pi, \zeta_i)$ is the loss associated with the policy π in relation to the specific preference ζ_i.

Compared to exact numerical rewards, preference feedback has several benefits (Wirth et al., 2017), such as (1) circumventing arbitrary reward design, reward shaping, reward engineering, or predefined objective trade-offs, (2) diminishing reliance on expert knowledge, and (3) decoupling training loop with human by modeling preferences (Akrour et al., 2012). However, PbRL also faces challenges, including credit assignment problems due to temporal delays, practical exploration of preference space (Wirth et al., 2017), the potential need for massive data (Ouyang et al., 2022), and the inability to use the learned preference model for retraining (McKinney et al., 2022).

Imitation Learning (IL) IL (Schaal, 1999; Syed et al., 2008), also referred to as learning from demonstration or apprenticeship learning, focuses on emulating human behaviors within specific tasks. The agent learns a mapping between observations and actions and refines its policy by observing demonstrations in a collection of teacher demonstration data $\mathcal{D}$ (Bakker et al., 1996; Hussein et al., 2017). This process obviates the need for environmental reward signals (Hussein et al., 2017). Broad IL (Cotra, 2018) aims to replicate human desires and intentions, effectively creating replicas of human decision-making processes. This concept is central to technologies such as Iterated Distillation and Amplification (IDA, as shown in §2.4.2) (Christiano et al., 2018). On the other hand, Narrow IL aims to replicate specific human behaviors within given tasks. Behavioral cloning (BC) (Bain and Sammut, 1995; Ross et al., 2011; Osa et al., 2018) is a simple (Pomerleau, 1991; Ravichandar et al., 2020) strategy that learns directly from demonstrations using supervised learning (Schaal, 1996). BC method specifically seeks to optimize the policy parameters, ϕ, with the objective of aligning the policy $\pi_\phi(a|s)$ closely with the expert policy $\pi_E(a|s)$. This alignment is achieved through the minimization of the negative log-likelihood, as delineated in the following (Lynch et al., 2020):

$$\mathcal{L}_{\text{BC}}(\phi) = -\mathbb{E}_{(s,a) \sim \pi_E} \left[\log \pi_\phi(a|s) \right].$$

Here, the expectation is computed over state-action pairs sampled from the expert policy, π_E. However, it faces the Out-of-Distribution (OOD) problem, arising from the difference between the training and testing distributions (Ross et al., 2011; Ho and Ermon, 2016; Reddy et al., 2019; Zhou et al., 2022). Adversarial imitation learning

---

[26]Notably, Sadigh et al. (2017) explicitly maintains a probabilistic belief over the true reward function during learning, and actively constructs queries to the human to reduce uncertainty maximally. Both traits are in a similar spirit to *cooperative inverse reinforcement learning* (CIRL), and later work also continues this theme (Reddy et al., 2020). See §2.4.5 for more.

methods (Ho and Ermon, 2016; Fu et al., 2018a; Lee et al., 2019; Ghasemipour et al., 2020) have demonstrated an ability to enhance the robustness of policies against distribution shifts. However, these methods learn non-stationary rewards, which cannot be used to train new policies (Ni et al., 2021).

Inverse Reinforcement Learning (IRL)   Unlike the paradigm of IL, IRL (Adams et al., 2022) focuses on deriving a reward function from observed behavior (Ng et al., 2000; Arora and Doshi, 2021). Standard IRL methods include the feature matching methods (Abbeel and Ng, 2004), which assumes optimal expert behavior or decision processes, as well as the maximum entropy methods (Ziebart et al., 2008) and the Bayesian methods (Ramachandran and Amir, 2007), both of which do not require optimal behavior. IRL guarantees robustness to changes in the state distribution but at the cost of increased computational complexity due to the extra RL step (Ho and Ermon, 2016; Fu et al., 2018b). This interaction, meanwhile, introduces inherent RL challenges, *e.g.*, sample efficiency (Yu, 2018) and potential dangers in environment interaction (García and Fernández, 2015). Additionally, identifying the reward function remains a challenge (Kim et al., 2021).

2.3.2   Reinforcement Learning from Human Feedback (RLHF)

RLHF expands upon PbRL within the domain of DRL (Christiano et al., 2017), aiming to more closely align complex AI systems with human preferences (OpenAI, 2023b). Its principal advantage is that it capitalizes on humans being better at judging appropriate behavior than giving demonstrations or manually setting rewards. This approach has gained significant traction, particularly in fine-tuning LLMs (Ouyang et al., 2022; OpenAI, 2023a; Touvron et al., 2023). Nonetheless, RLHF encounters obstacles (Casper et al., 2023b), including data quality concerns, the risk of reward misgeneralization, reward hacking, and complications in policy optimization. Specifically, RLHF can also be viewed as a Recursive Reward Modeling (RRM) process (as shown in §2.4.3) without deep recursive modeling (Leike et al., 2018). Here, we provide a brief review of the RLHF methodology.

The genesis of RLHF can be traced back to Knox and Stone (2008, 2012), subsequently broadening its reach to domains such as social robots (Knox et al., 2013) and human-AI cooperative learning (Griffith et al., 2013). Besides focusing on the association between feedback and policy, Loftin et al. (2016) models the connection between feedback and the trainer strategy. Christiano et al. (2017) extended RLHF to simulated robotic tasks, demonstrating its potential effectiveness.

It's worth noting that one of the significant applications of RLHF has been in the field of LLMs. Some work found that LLMs trained with RLHF (Ouyang et al., 2022; Korbak et al., 2023; Christiano, 2023) are more creative and human alignment compared to models trained via supervised or self-supervised learning approaches (Kenton and Toutanova, 2019; Brown et al., 2020b). The importance of RLHF is not merely limited to allowing LLMs to follow human directives (Ouyang et al., 2022). It helps LLMs better align by giving them important qualities like being helpful, harmless, and honest (Bai et al., 2022a). Due to these improvements, many works use RLHF for aligning LLMs (Ziegler et al., 2019; Stiennon et al., 2020; Bai et al., 2022a; Glaese et al., 2022; OpenAI, 2023a; Touvron et al., 2023). Additionally, Dai et al. (2024) integrates the Safe RL (García and Fernández, 2015) framework with the RLHF, addressing the inherent tension between aligning helpfulness and harmfulness (Bai et al., 2022a). Future efforts can be focused on reducing dependence on human annotation (Wang et al., 2023c; Sun et al., 2024) and improving the efficacy of the reward model by leveraging iterative RLHF methods (*i.e.*, integrating it with debate frameworks (Irving et al., 2018)), *etc.* Qiu et al. (2024) has also built a formal framework of the RLHF process portraying it as an autoencoding process over text distributions, and enables analysis of convergence properties in RLHF.

We review the RLHF pipeline from the Ziegler et al. (2019); Ouyang et al. (2022); Rafailov et al. (2024) to give a general framework. It usually consists of three stages:

- **Supervised Fine-tuning (SFT)**. RLHF usually starts with a pre-trained language model, then fine-tuned using supervised learning – specifically, maximum likelihood estimation – on a high-quality human instruction dataset tailored for downstream tasks to obtain a model π^{SFT}. Examples of these tasks include dialogue handling, instruction following, and summarization (Some open-source datasets include Alpaca Data (52k instruction-following data) (Taori et al., 2023), Vicuna (70K user-shared ChatGPT conversations) (Chiang et al., 2023), *etc.*). This stage can also be carried out at any other stage.

- **Collecting Comparison Data and Reward Modeling**. This phase includes collecting comparison data, which is subsequently used to train a reward model. The SFT model is given prompts denoted as x to generate pairs of responses (y_1, y_2) sampled from $\pi^{\text{SFT}}(y \mid x)$. These pairs are subsequently shown to human annotators, who indicate a preference for one of the responses. Then as discussed in §2.2, comparison data is used to construct the reward model r_θ.

- **Policy Optimization via Reinforcement Learning**. The final step is optimizing LLM as a policy π through RL, guided by the reward model r_θ. The process of LLMs generating responses from prompts is modeled as

a bandit environment (Ouyang et al., 2022), where a reward is obtained from reward model r_θ at the end of each response. The primary objective of RL is to adjust the parameters ϕ of the LLMs such that the expected reward on training prompt dataset $\mathcal{D}_{\text{RL}}$ is maximized:

$$\arg\max_{\pi_\phi} \mathbb{E}_{x \sim \mathcal{D}_{\text{RL}}, y \sim \pi_\phi}\Big[r_\theta(x, y)\Big].$$

Typically, an additional per-token KL penalty derived from the SFT model π^{SFT} is involved to mitigate the reward over-optimization. In addition, the integration of gradients from pre-training distribution $\mathcal{D}_{\text{pretrain}}$ helps maintain model performance, denoted as PTX loss in (Ouyang et al., 2022). As a result, a more comprehensive practical objective function is introduced:

$$\mathcal{J}(\phi) = \mathbb{E}_{x \sim \mathcal{D}_{\text{RL}}, y \sim \pi_\phi}\Big[r_\theta(x, y) - \beta \log\big(\pi_\phi(y|x)/\pi^{\text{SFT}}(y|x)\big)\Big] + \eta\, \mathbb{E}_{(x,y) \sim \mathcal{D}_{\text{pretrain}}}\Big[\log\big(\pi_\phi(y|x)\big)\Big],$$

where β and η are coefficients determining the intensity of the KL penalty and the mixture of pretraining gradients respectively. This process refines the LLMs to generate responses that better align with human preferences for the prompts used during training.

Though RLHF has been proven effective for aligning LLMs with human preferences, this method has problems like complex implementation, hyper-parameter tuning, sample efficiency (Choshen et al., 2019), and computational overhead (Yuan et al., 2024), making it hard to scale up.

A straightforward approach is rejection sampling (Dong et al., 2023; Touvron et al., 2023) paired with finetuning on the best examples. For every prompt, K responses are sampled from the model. Each response is then assessed with the reward model, and the one with the highest reward is selected as the best response. This selected response is later used for model fine-tuning. Zhang et al. (2023a) formulates the language model instruction alignment problem as a goal-reaching reinforcement learning problem and proposes the HIR algorithm. The method unfolds in two stages: online sampling and offline training. During online sampling, the algorithm samples the LLM at a high temperature. In the offline training stage, instructions are relabeled based on generated outputs, followed by supervised learning using this relabeled data. HIR capitalizes on successful and failed cases without requiring additional parameters. RRHF, as introduced by (Yuan et al., 2024), aligns model probabilities with human preferences by scoring and ranking responses from multiple sources. With the necessity for only 1 or 2 models, its implementation is straightforward. RRHF reported it can effectively align language models with human preferences, producing performance on par with PPO. Gulcehre et al. (2023) proposes the ReST algorithm, which contains two loops: *Grow* and *Improve*. The *Grow* loop uses the current model to sample and generate a dataset, while the *Improve* loop iteratively trains the model on a fixed dataset. This algorithm provides a simple and efficient framework that allows repeated use of the fixed dataset to improve computational efficiency, showing significant improvement in the reward model scores and translation quality compared to supervised learning baselines. Motivated by the dependence of reward modeling on policy optimization in RLHF, Chakraborty et al. (2024) propose PARL, a bilevel optimization-based framework.

Rafailov et al. (2024) introduces the DPO, which demonstrates a mapping between reward functions and optimal policies. DPO is both simple and efficient, optimizing language models directly from human preference data, eliminating the need for an explicit reward model and multi-stage training. Azar et al. (2023) presents a general objective, ΨPO, designed for learning from pairwise human preferences, circumventing current methods' assumption: *pairwise preferences can be substituted with pointwise rewards*. This objective analyzes RLHF and DPO behaviors, revealing their potential overfitting issue. The authors further delve into a specific instance of ΨPO by setting Ψ as the Identity, aiming to mitigate the overfitting problems. They call this method IPO and furnish empirical results contrasting IPO with DPO. Hejna et al. (2024) introduces CPL, which utilizes a regret-based model of preferences that directly provides information about the optimal policy.

Further research could explore why RLHF performs effectively with LLMs and the application of RLHF in multimodal (Yevgen Chebotar, 2023; OpenAI, 2023b) settings to facilitate the benefits of human-AI collaboration (Carlson and Demiris, 2010; Wu et al., 2021; Bi et al., 2021). See also Casper et al. (2023b) who offer a survey of open problems with RLHF.

2.4 Scalable Oversight

Statistical learning algorithms usually rely on certain assumptions about data distribution, such as independence and identical distribution. Consequently, these algorithms fail in some situations, especially under specific distributions (Zhou et al., 2022). Challenges in elementary systems can be promptly identified through visual inspection (Christiano et al., 2018; Ngo et al., 2024). As AI systems become more powerful, insufficiently capturing the training signal or erroneous design of loss functions often leads to catastrophic behaviors (Russell et al., 2015; Hubinger
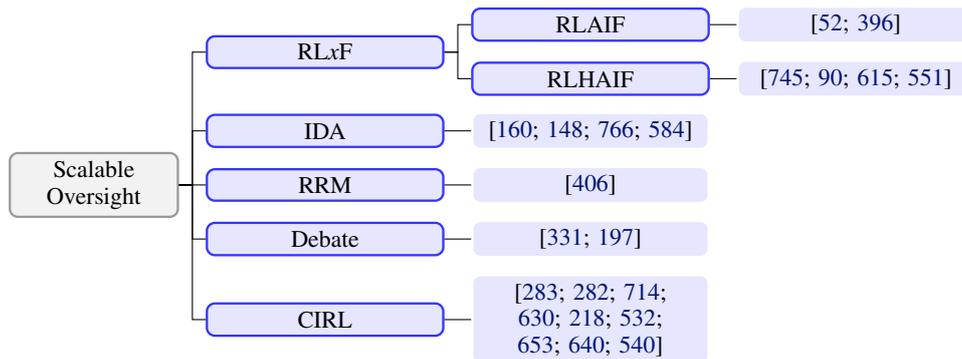
Figure 5: A tree diagram summarizing the key concepts and literature related to Scalable Oversight. The root node represents Scalable Oversight whose goal is *ensuring AI systems remain aligned with human intent even as they surpass human capabilities*. The main branches represent promising frameworks such as Reinforcement Learning from Feedback (RLxF), Iterated Distillation and Amplification (IDA), Recursive Reward Modeling (RRM), Debate, and Cooperative Inverse Reinforcement Learning (CIRL). Further sub-branches list key works exploring each framework. This diagram provides an overview of research directions for constructing effective and safe oversight mechanisms as AI systems grow more complex.

et al., 2019c; Cotra, 2021) such as deceiving humans by obfuscating discrepancies (Russell, 2019), specification gaming (Victoria et al., 2020), reward hacking (Brown et al., 2020a), and power-seeking dynamics (Carlsmith, 2022). From a human perspective, these imply gaps between the optimized objectives of AI systems and the ideal goals in our minds. Thus, the issue of providing effective oversight in various decision-making becomes pivotal (Bowman et al., 2022; Li et al., 2023a), often termed as *scalable oversight* (Amodei et al., 2016) arising from two practical challenges.

- The high cost of humans frequently evaluating AI system behavior. For instance, the training process is time-consuming, and incorporating humans directly into the training loop in real-time would significantly waste human resources and impede training efficiency (Christiano et al., 2017).

- The inherent complexity of AI system behaviors makes evaluation difficult, especially on hard-to-comprehend and high-stakes tasks (Saunders et al., 2022), *e.g.*, tasks such as teaching an AI system to summarize books (Wu et al., 2021), generate complex pieces of code (Pearce et al., 2022), and predict future weather changes (Bi et al., 2023).

> *Scalable oversight seeks to ensure that AI systems, even those surpassing human expertise, remain aligned with human intent.*

In this context, our primary focus is to present some promising directions that may have not yet been implemented generally for constructing scalable oversight (Amodei et al., 2016; Leike et al., 2018).

2.4.1   From RLHF to RLxF

The RLHF paradigm offers a framework for aligning complex systems (OpenAI, 2023a; Touvron et al., 2023). However, it encounters obstacles such as the inaccuracy of human evaluations and their associated high costs (Christiano et al., 2017; Casper et al., 2023b; Perez et al., 2023). A key limitation is the difficulty in utilizing RLHF to extend human feedback when creating AI systems with superhuman abilities (Wu et al., 2021). Building on the RLHF paradigm, we introduce *RLxF* as a fundamental framework for scalable oversight, aiming to enhance feedback efficiency and quality and expand human feedback for more complex tasks. This enhances RLHF by incorporating AI components (Fernandes et al., 2023). The *x* in *RLxF* signifies a blend of AI and humans. We further explore concrete methodologies about *RLxF* in the subsequent section.

Reinforcement Learning from AI Feedback (RLAIF)   RLAIF is a method building upon the framework of RLHF and serves as an extension to RLHF. Bai et al. (2022a) found that LLMs trained via RLHF often select to avoid sensitive and contentious issues, potentially diminishing models' overall utility. Considering these limitations, Bai et al. (2022b) proposed a training pipeline based on RLAIF, which uses feedback generated by the LLMs (*e.g.*, GPT-4 or other language models having superhuman capabilities) rather than human feedback. Based on pre-set criteria, the policy model self-evaluates and revises its responses prompted by *red teaming*. Then, they

fine-tuned the initial policy model with revised responses. Finally, the fine-tuned policy model assesses harmlessness for another language model's response (*i.e.*, AI feedback). Mirroring the RLHF method, they train a reward model using this feedback and optimize the behavior of the policy model. Lee et al. (2023a) compares the performance differences between models trained with RLAIF and RLHF on the summarization task. Their results suggest that models trained with AI feedback achieved nearly identical overall performance to those trained with human feedback when evaluated by humans, though there are nuances.

To some extent, RLAIF addresses the evasion (Bai et al., 2022b) inherent in RLHF (*i.e.*, keep harmlessness without appreciable utility decline). AI Feedback offers a viable alternative for constructing a training loop that necessitates minimal human intervention, reducing the cost of training. AI supervision obeying transparent and accessible AI behavior guidelines may significantly aid in achieving scalable oversight (Bowman et al., 2022).

Reinforcement Learning from Human and AI Feedback (RLHAIF)   RLHAIF integrates human and AI elements to provide oversight. Wu et al. (2021) investigates the feasibility of AI in assisting humans in summarizing books. This method facilitated human supervision and evaluation of the model's performance by decomposing the book summarization task into subtasks to form a tree-like structure. Meanwhile, Saunders et al. (2022) explores the feasibility of leveraging AI to aid in the human assessment of model efficacy. Their findings indicate that model-generated critiques help humans identify flaws they may have missed. Bowman et al. (2022) proposes a proof-of-concept experiment to demonstrate the promising to evaluate scalable oversight techniques based on *sandwiching* (Cotra, 2021). When collaborating with an unreliable LLM, the outcomes reveal that humans significantly surpass the model and themselves. Perez et al. (2023) employs language models to autonomously generate datasets for evaluating the behavior of language models of varying scales. The authors produced 154 high-quality datasets validated by humans. These methods demonstrate the feasibility of using AI assistance to scale up human oversight over complex problems and various domains.

Discussion   Some efforts are underway to enhance RLHF algorithms by replacing pure humans with other components (Leike et al., 2018). Given the multidimensional nature of human feedback, various approaches have been devised to offer focused human judgments informed by specific rules. Examples of such rules encompass considerations like chat fluency (Saunders et al., 2022) and privacy safeguards (Carr, 2023). Saunders et al. (2022) deconstructs the requirements for quality dialogue into natural language guidelines that an agent should adhere to, asking for evaluations on each guideline individually. We can attain more efficient rule-conditioned reward models by collecting targeted human assessments and training models on this data. This approach substantially enhances the efficacy of dialogue agents, rendering them more helpful, accurate, and benign when compared to prompted language models. Carr (2023) proposes Reinforcement Learning from Privacy Feedback (RLPF), aiming to harmonize the output quality of language models with safeguarding privacy. The method exploits NLP techniques to conduct real-time privacy risk assessments of text generated by the models and subsequently adjusts the reinforcement learning feedback signals based on these evaluations. Expressly, if the generated text includes sensitive information, it incurs negative feedback, whereas high-quality, non-revelatory text receives positive feedback. As the model undergoes training, it incrementally refines its capabilities, enhancing text quality and minimizing privacy breaches concurrently. This approach offers a more efficient evaluation of privacy risks by employing established NLP techniques, in contrast to conventional learning methods, which depend heavily on large-scale manual data annotation.

At their core, the *RLxF* methods utilize the strategy of decomposing a large problem into smaller sub-problems, enabling the use of more efficient tools, such as AI and software, for rapid sub-problem resolution. By leveraging the solutions to these sub-problems, the resolution of the main issue can be expedited. These techniques can be regarded as elementary instances of IDA; the primary distinction lies in the absence of a continual iterative process. Nonetheless, evidence suggests they are promising to offer feedback for AI systems that exceed human performance (Wu et al., 2021). Consequently, these methods can serve as foundational techniques in the training of more advanced AI systems.

2.4.2   Iterated Distillation and Amplification

Iterated Distillation and Amplification (IDA) introduces a framework for constructing scalable oversight through iterative collaboration between humans and AIs (Christiano et al., 2018). The process commences with an initial agent, denoted as $A[0]$, which mirrors the decision-making of a human, H. $A[0]$ undergoes training using a potent technique that equips it with near-human-level proficiency (the distillation step); Then, collaborative interaction between H and multiple $A[0]$ instances leads to the creation of an enhanced agent, $A[1]$ (the amplification step). The successive process is described[27] in Algorithm 1.

Cotra (2018) distinguishes between broad and narrow definitions within both RL and IRL. Broad RL gives sparse reward signals to AI systems and allows autonomous exploration and optimization of cumulative future

---

[27]We reference the pseudo-code by Cotra (2018) for this description.

Algorithm 1 Iterative Distillation and Amplification

 1: **procedure** IDA(H)
 2:     $A \leftarrow$ random initialization
 3:     **repeat**
 4:         $B \leftarrow$ AMPLIFY(H, A)
 5:         $A \leftarrow$ DISTILL(B)   ▷ Repeat indefinitely
 6:     **until** False
 7: **end procedure**
 8: **procedure** DISTILL(overseer)
        return An AI trained using narrow, robust techniques to perform a task that the overseer already understands how to perform.
 9: **end procedure**
10: **procedure** AMPLIFY(human, AI)
        ▷ Interactive process in which human uses many calls to AI to improve on human's native performance at the relevant tasks.
11: **end procedure**

---

rewards. This can lead to super-human novel strategies but makes it hard to specify what we care about perfectly. Narrow RL gives dense feedback rewarding the reasonableness of choices instead of final outcomes. This makes ML systems more human-like but limits capabilities. Similarly, broad IRL infers deep long-term values from the full range of human behaviors, while narrow IRL only infers short-term instrumental values. The former is a higher risk, while the latter is limited in capabilities.

During IDA training, narrow techniques are needed to ensure each agent itself mimics human behaviors. Specifically, narrow RL or IL can be used to train the agent to be as human-like and controllable as possible. Humans can leverage agents' computing power and parallelizability to devise more far-sighted, macro strategies. This is essentially an amplification of human intrinsic capabilities. In the next iteration, agents again mimic this strengthened human-machine system using narrow techniques. This enables a gradual transition from narrow ability to broad ability while keeping the agents aligned with human values. As iterations increase, the human-machine system becomes more and more capable, gradually approximating a system that is both highly capable and aligned with human values, achieving both safety and capability. In other words, Narrow techniques are used to ensure agents follow human values, while the broadened human strategies in the amplification stage are a way of utilizing the agents, and do not expand the agents' own learning goals.

IDA is well illustrated by AlphaZero (Christiano et al., 2018; Nguyen, 2020). The algorithm starts with a simple policy (*e.g.*, random move selection) and learns from its self-play games, the *amplification* phase. It then uses these games as training data to develop better move selection heuristics, the *distillation* phase. This distillation-amplification process can be repeated to create a fast and proficient Go-playing AI. Here, the distinction between alignment and capability is crucial (Mennen, 2018). An aligned but less capable AI tries to win but may not succeed against moderate opponents. A capable but poorly aligned AI achieves certain game properties other than winning. The goal is that AI is capable and aligned, proficient at the game, and aligned with the goal of winning the game.

The feasibility of IDA has sparked considerable debate (Yudkowsky, 2018). IDA operates under a crucial assumption that *errors won't continuously accumulate throughout the iterations* (Leike et al., 2018). Thus, technical challenges persist during the distillation and amplification step, necessitating sufficiently advanced and safe learning techniques. Additionally, despite the original authors likening IDA to the training process of AlphaZero (Silver et al., 2017) and having demonstrated it in toy environments (Christiano et al., 2018), its practicality hinges on ensuring that H can delegate portions of complex tasks to A, analogous to a leader orchestrating a team to accomplish a project collectively. In practice, Gato (Reed et al., 2022) illustrates key aspects of IDA (Mukobi, 2022) that may pave the way to AGI. It consolidates the abilities of multiple expert AIs into a singular model, validating that IDA's distillation can be achieved using contemporary deep learning. While not fully realized, Gato hints at amplification potential, harnessing its diverse skills to accelerate the learning of new tasks. However, Gato lacks safe amplification or distillation methods to maintain alignment properties. Crafting alignment-preserving IDA methods suited for models like Gato remains a crucial direction for AI safety research. In essence, while Gato signifies notable progress in actualizing IDA, further theoretical advancements are imperative to ensure that the IDA framework leads to safe AGI.

Algorithm 2 Recursive Reward Modeling

1: Initialize agent A_0 using reward modeling based on user feedback. ▷ Either preferences or numerical signals.
2: **for** $t = 1, 2, \ldots$ **do**
3:     Use A_{t-1} to assist users in evaluating outcomes.
4:     Train agent A_t based on user-assisted evaluations.   ▷ Objective of A_t is generally more complex than that of A_{t-1}.
5: **end for**

2.4.3   Recursive Reward Modeling

As discussed in §2.2, reward modeling leverages the idea of using human feedback to train a reward model, which an agent then pursues. It allows us to disentangle the construction of the system's objective from evaluating its behavior (Ibarz et al., 2018). In this manner, the reward model provides insights into the optimization direction of the AI system. Particularly noteworthy is the ability to finely align the system with human intentions and values, such as fine-tuning language models to adhere to human instructions (Bai et al., 2022a; Touvron et al., 2023). Also, reward modeling has proved valuable in advancing AI research (Zhao et al., 2023; Bukharin et al., 2023). Recursive Reward Modeling (RRM) (Leike et al., 2018) seeks to broaden the application of reward modeling to much more intricate tasks. The central insight of RRM is the recursive use of already trained agents A_{t-1} to provide feedback by performing reward learning on an amplified version of itself for the training of successive agents A_t on more complex tasks. The A_0 is trained via fundamental reward modeling (learned from pure human feedback). This approach is not only influenced by human feedback but also by the model's own assessments of what constitutes a rewarding outcome. If the assumption that *evaluating outcomes is easier than producing behavior* holds, then the iterative process of reward modeling can iteratively achieve higher capacity to oversee more powerful AI systems, paving the way for extending oversight into more complex domains. This process is detailed in Algorithm 2.

For instance, we aim to train AI A to devise a comprehensive urban plan. Designing a city entails numerous intricate elements, such as traffic planning, public amenities, and the distribution of residential and commercial zones. Evaluating a city's entire design poses a significant challenge since many issues may only become apparent after extended real-world testing. To aid this process, we may need an agent B specifically for traffic planning. However, traffic planning in itself is a multifaceted task. Consequently, we further need other agents to assess aspects such as road width, traffic flow, and the design of public transportation. For every sub-task, such as gauging road width, we can train an auxiliary agent to verify if safety standards are met, if various modes of transportation have been considered, and so on. In doing so, we establish an RRM process where each agent is trained with the help of agents assessing sub-tasks.

This approach resembles the organizational structure of a large corporation (Leike et al., 2018). In the context of urban planning, the main planning team (the CEO) is responsible for the final design decisions. Their decisions are informed by recommendations from the traffic team (the department managers), who, in turn, base their recommendations on inputs from the road width team (the managers), and so forth. Each level of decision-making relies on feedback from the level below it, with each task optimized through reward modeling.

The challenges faced by RRM can be described around the concepts of outer and inner alignment (Hubinger, 2020). Outer alignment revolves around the sufficiency of feedback mechanisms to guarantee that the learned reward model is accurate in the domain perceived by the action model as on distribution. This challenge is contingent on several factors, including the quality of human feedback, the difficulty of generalization, and the potential for agent deception. Conversely, inner alignment concentrates on how effectively a human can employ transparency tools to prevent deceptive or disastrous behaviors in both the reward model and the agent. This hinges on the effectiveness of the oversight mechanism and the capacity to verify that the reward model isn't undergoing any optimization and that the agent remains myopic (Cotra, 2018).

Potential approaches to mitigate these challenges (Leike et al., 2018) include online feedback to correct the reward model during training (Christiano et al., 2017), off-policy feedback to teach about unsafe states (Everitt et al., 2017), leveraging existing data like videos and text via unsupervised learning or annotating (Baker et al., 2022), hierarchical feedback on different levels (Bukharin et al., 2023) adversarial training to discover vulnerabilities (Madry et al., 2018), and uncertainty estimates for soliciting feedback (Hadfield-Menell et al., 2016; MacGlashan et al., 2017). The strength of RRM is its competitive training approach, which necessitates human feedback instead of demonstrations, potentially making feedback more reliable and simpler to obtain (Hubinger, 2020). In essence, the process of RRM can be likened to IDA (Christiano et al., 2018), where reward modeling takes the place of supervised or imitation learning. Thus, the challenges confronted by RRM closely mirror those encountered in IDA, particularly in preventing the accumulation of errors. Additionally, reward modeling itself does not necessarily distill a *narrow* model (Cotra, 2018), which presents challenges in trading off the degree of alignment and performance.

Algorithm 3 Debate

1: Initialize set of questions Q.
2: Initialize two competing agents.
3: Select a question $q \in Q$.   ▷ Question is shown to both agents.
4: Agents provide their answers a_0 and a_1. The agents generate comment answers in response to q.
5: Initialize debate transcript T as an empty list.
6: **for** turn in predefined number of debate turns **do**
7:       Agent makes a debate statement s.
8:       Append s to T.   ▷ Agents take turns and statements are saved in the transcript.
9: **end for**
10: Judge observes (q, a_0, a_1, T) and decides the winning agent.

2.4.4   Debate

Debate involves two agents presenting answers and statements to assist human judges in their decision-making (Irving et al., 2018), as delineated in Algorithm 3. This is a zero-sum debate game where agents try to identify each other's shortcomings while striving to gain higher trust from human judges, and it can be a potential approach to constructing scalable oversight. For example, in the game of Go, human judges might not discern the advantage side of the single game board itself. However, by observing the game's process and the eventual outcome, these judges can more easily deduce that.

The premise of this method relies on a critical assumption: *arguing for truth is generally easier than for false-hood*, granting an advantage to the truth-telling debater. However, this assumption does not hold universally. For instance, in a complex problem, humans might fail to comprehend the specialized concepts used in the debate. Additionally, the limited nature of the gradient descent may bring us to an undesirable cyclic pattern (*i.e.*, when optimizing for one property, such as honesty and highlighting flaws, models often overlook or diminish another) (Irving et al., 2018).

It's worth mentioning that with the advancement of LLMs' capabilities, we can already see practical examples of debate (Du et al., 2023; Claude, 2023). Challenges may arise for debate in specific real-world scenarios (Irving et al., 2018). For example, certain questions may be too intricate for human comprehension or too voluminous to present in their entirety. Consider the complexity of interpreting a 10-gigapixel image or sifting through the vastness of the entire internet. Similarly, there are instances where an optimal answer to a question is exceedingly lengthy. Envision needs a response that spans a hundred pages. To navigate these, agents might initially select a response and, as the debate progresses, reveal sections of either the question or the answer. Irving et al. (2018) conducts a toy experiment on this process. Meanwhile, we must acknowledge the limit of human time. In scenarios that necessitate interaction with the environment, such as directing a robot, each action might demand a distinct debate. It's not always feasible for humans to judge every debate due to time constraints. In response to this challenge, we may need to design ML models to predict human feedback.

Another consideration is the convergence of the debate mechanism (Irving et al., 2018). Du et al. (2023) show-cases the inherent tendency of the debate framework to eventually converge toward singular responses, even if accuracy is not guaranteed. Meanwhile, if challenges arise in achieving convergence, we might have to rely on intuition to gauge the effectiveness of convergence. This implies the requirement of human evaluators' intervention and demands a certain level of expertise from these human assessors, posing challenges that must be addressed.

Furthermore, there are many discussions originating from diverse perspectives. Ngo (2021) considers *Debate* as one type of iterated amplification but more specific to make safety ground in concrete research questions, and its adversarial framing makes it easier to spot problems. Michaelcohen (2020) expresses concerns regarding the adverse implications of incentivizing debaters to employ deceptive strategies aimed at swaying the judgment process. Armstrong (2019); Barnes (2020) expound upon the various issues that can permeate the debate process, including challenges such as the obfuscated arguments problem, ambiguous responses, and the propagation of misleading implications. While one may affirm the presence of a sufficiently low probability of any underlying flaws within the argument, advocating for trustworthiness, the opposing debater may assert the existence of a sufficiently high probability of identifying a flaw within the argument somewhere, thus advocating for a lack of trust. Beth Barnes (2020) introduces the concept of *cross-examination* to incentivize debaters to provide more informative responses. In this process, debaters have the agency to select a prior claim for scrutiny and obtain a copy of the opposing debater's response. The entire exchange is documented, and debaters can present relevant segments to the judge. The introduction of cross-examination is a robust deterrent against dishonest debaters exploiting a sweeping narrative, in contrast to their prior arguments, to mislead the judge.

There exists a notable similarity between the debate (Irving et al., 2018), IDA (Christiano et al., 2018), and RRM (Leike et al., 2018). These approaches can be comprehended in the view of an underlying principle: *evaluation*

can be simpler than task completion[28]. Therefore, harnessing the evaluative capabilities of AI systems can result in distributions of capacity that are more advantageous for humans. The challenges these methods face, especially in mitigating the accumulation of errors, are also analogous.

2.4.5   Cooperative Inverse Reinforcement Learning

Almost all previous methods consider learning from feedback a process separate from inference and control and often implicitly consider feedback providers as entities existing outside of the environment – indeed, failure modes like manipulation (Shevlane et al., 2023) and reward tampering (Everitt et al., 2021) occur exactly when feedback mechanisms that are supposedly outside of the environment become part of it and therefore subject to the AI system's influence. The framework of Cooperative Inverse Reinforcement Learning (CIRL), however, unifies control and learning from feedback and models human feedback providers as fellow agents in the same environment. It approaches the scalable oversight problem not by strengthening oversight but by trying to eliminate the incentives for AI systems to game oversight, putting humans giving feedback and the AI system in cooperative rather than adversarial positions (Shah et al., 2020). In the CIRL paradigm, the AI system collaborates with humans to achieve the human's true goal rather than unilaterally optimizing for human preferences.

Motivation and General Idea of CIRL   Many modes of misalignment, including, for example, reward hacking (Victoria et al., 2020; Skalse et al., 2022), deception (Park et al., 2023b), and manipulation (Shevlane et al., 2023), are results of the AI system confidently optimizing for misspecified objectives (Pan et al., 2021). During training and deployment, the specified objective (*e.g.*, the reward function) plays the role of an unchallengeable truth for the AI system, and human feedback is only respected to the extent specified in the objective, which means that it could be tampered (Everitt et al., 2021) or manipulated (Shevlane et al., 2023).

CIRL (Hadfield-Menell et al., 2016, 2017b; Shah et al., 2020) attempts to mitigate this problem by (1) having the AI system explicitly hold uncertainty regarding its reward function, and (2) having humans provide the only information about what the reward function truly is. This uncertainty gives the AI system a tendency to defer to humans and a drive to determine what the human truly wants. Concretely speaking, it models the entire task as a two-player cooperative game, where the *human player H* and the *robot player R* share a common reward function $r(\cdot)$. Importantly, the reward function and reward signals aren't visible to R (and indeed aren't explicitly calculated by the training mechanism) and are only inferred by R from behaviors of H via an IRL-like process (including by asking and interacting with H). This game has been called the *CIRL* (Hadfield-Menell et al., 2016), the *assistance game* (Fickinger et al., 2020), and the *assistance POMDP* (Shah et al., 2020).

In short, the AI system has the human's true objective $r(\cdot)$ as its own goal (despite not knowing values of $r(\cdot)$ with certainty) and constantly tries to figure r out by observing and interacting with the human. This reduces incentives for, *e.g.*, manipulation since manipulation of human behaviors only serves to pollute an information source and does not affect r.

Formulation of CIRL   Hadfield-Menell et al. (2016) characterizes the settings of CIRL (which we denote by M) by building upon classical multi-agent MDPs, resulting in the definition below of M.

$$M = \left\langle \mathcal{S}, \{\mathcal{A}^{\mathbf{H}}, \mathcal{A}^{\mathbf{R}}\}, T, \gamma, r, \Theta, P_0 \right\rangle$$

In the equation above, S and $\{\mathcal{A}^{\mathbf{H}}, \mathcal{A}^{\mathbf{R}}\}$ are the space of world states and actions respectively, $T : S \times \mathcal{A}^{\mathbf{H}} \times \mathcal{A}^{\mathbf{R}} \to \Delta(S)$ is the transition function, and γ is the discount rate. Up to here, the definition is identical to that of a standard multi-agent MDP. The remaining elements, however, introduce the key difference: the reward function is parameterized, and its parameters can be modeled by a distribution. Θ is the space of values for the parameters θ; $r : S \times \mathcal{A}^{\mathbf{H}} \times \mathcal{A}^{\mathbf{R}} \times \Theta \to \mathbb{R}$ is the shared reward function, and $P_0 \in \Delta(S \times \Theta)$ is the joint distribution of the initial state and the reward function's parameters. This parameterization approach allows R to model explicitly and reason about its belief over the true reward function. Using techniques from Nayyar et al. (2013), any CIRL setting can be reduced to an equivalent single-agent POMDP, thus proving the existence of optimal policies that are relatively tractable (Hadfield-Menell et al., 2016).

Notable Directions in CIRL Research   Although some have emphasized the importance of H teaching R (Fisac et al., 2020) actively, works (Shah et al., 2020) have contested the emphasis on game equilibria and joint policies (including H's pedagogic behaviors), and instead focuses on R's optimal response to a policy of H's, since the assumption that humans will always act on optimal joint policies is an unrealistic one. More specifically, Shah et al. (2020) considers the *policy-conditioned belief* $B : \Pi^{\mathbf{R}} \to \Delta\left(\Pi^{\mathbf{H}}\right)$, which specifies H's distribution over policy responses to any of R's policies, and the aim is to find R's optimal policy given B. Here, B is essentially a form of human modeling, and one challenge is to obtain a robustly accurate human model as B (Hong et al., 2022).

---

[28]Discussions about this can also be found in the literature about these methods.

On another front, Hadfield-Menell et al. (2017b) and He and Dragan (2021) examine the manual specification of an imperfect reward function as a way for H to convey information about the true reward function. This includes work on R's side (*i.e.*, enabling R to perform inference on the true reward function based on the imperfect specification) (Hadfield-Menell et al., 2017b) and also work on H's side (*i.e.*, developing algorithmic tools to assist H in making more robust specifications that better convey the true reward function) (He and Dragan, 2021). Aside from improvements to the game settings, the design of more scalable CIRL algorithms has also been recognized as a priority.

There has also been work that extends CIRL and assistant games to multi-agent settings (Fickinger et al., 2020) where there are multiple humans that the robot needs to serve. This corresponds to the *multi/single delegation* settings in Critch and Krueger (2020), where the varying objectives of humans create a challenge and necessitate the use of social choice methods.

2.5   Weak-to-Strong Generalization

Scalable Oversight can help humans provide supervision signals to AI systems that are smarter and more complex, ensuring that the behaviors of super-human-level AI systems align with human intent and values. However, what if we cannot obtain scalable supervision signals? An example is that for some tasks, evaluation is not necessarily simpler than generation, making it impossible to utilize task decomposition followed by AI assistance to achieve scalable oversight.

Recently, a generalization phenomenon called *Weak-to-Strong Generalization* is verified, the core idea of which is to use weak supervision signals from a weak model to train a strong model (Burns et al., 2023). Specifically, the weak model is trained on ground truth and then annotates new data with weak labels for training the strong model. The results across three settings (*i.e.* NLP classification, chess puzzles and reward modeling) reflect that *weak-to-strong generalization* is a robust phenomenon, yet there is room for further improvement, such as narrowing the gap between a strong model trained with weak labels and ground truth. *Weak-to-Strong Generalization* provides a valuable analogy for the superalignment problem: how humans can supervise super AI systems as weak supervisors. The insight behind *weak-to-strong generalization* is that the strong model can generalize beyond weak labels instead of merely imitating the behavior of weak models. In other words, the weak model elicits the strong model's capability. However, verifying *weak-to-strong generalization* is challenging if humans don't know the ground truth. Nonetheless, *weak-to-strong generalization* still offers a valuable perspective for solving the superalignment problem.

The framework for *weak-to-strong generalization* has been further expanding and integrating with scalable oversight. Empirical results show that weak models can evaluate the correctness of stronger models by assessing the debate between two expert models (Khan et al., 2024). Additionally, making expert debaters more persuasive improves non-experts' ability to discern truth in debates, evidencing the effectiveness of aligning models with debate strategies without ground truth. Some frameworks employ a external amplifier to create an iterated distillation and amplification process, which presents a potential framework for integrating *weak-to-strong generalization* techniques with IDA during the training process (Ji et al., 2024a). Moreover, Leike (2023a) proposes several methods to integrate scalable oversight with *weak-to-strong generalization* techniques, *e.g.*, recursively decomposing tasks into atomic ones (in line with scalable oversight principles), supervising these atomic tasks, and employing reward models trained with *weak-to-strong generalization techniques* using human preference data.

3   Learning under Distribution Shift

The construction of reliable AI systems is heavily dependent on their ability to adapt to diverse data distributions. Training data and training environments are often imperfect approximations of real deployment scenarios and may lack critical elements such as adversarial pressures (Poursaeed et al., 2021) (*e.g.*, Gaussian noise in the context of supervise learning-based systems (Gilmer et al., 2019) and shadow attack (Ma et al., 2012) in autonomous-driving systems), multi-agent interactions (Critch and Krueger, 2020; Dafoe et al., 2021), complicated tasks that human overseers cannot efficiently evaluate (Leike et al., 2018),[29] and reward mechanisms that can be gamed or manipulated (Krueger et al., 2020). This discrepancy between training distribution and testing distribution (or environments) is known as *distribution shift* (Krueger et al., 2020; Thulasidasan et al., 2021).

Therefore, AI systems that are aligned under their training distribution (*i.e.*, pursuing goals that are in line with human intent) may not uphold their alignment under deployment (or testing) distribution, potentially leading to serious misalignment issues post-deployment. This potential failure motivates research on the preservation of alignment properties (*i.e.*, adherence to human intentions and values) across data distributions.

From an alignment perspective, we are more concerned about AI systems pursuing unaligned and harmful goals, as opposed to incompetence at pursuing goals. Thus, the emphasis on alignment properties means that we focus on

---

[29]This could contribute to the emergence of deceptive behaviors (Hubinger et al., 2019a). See the paragraph on *goal misgeneralization* in §3.1 for details.
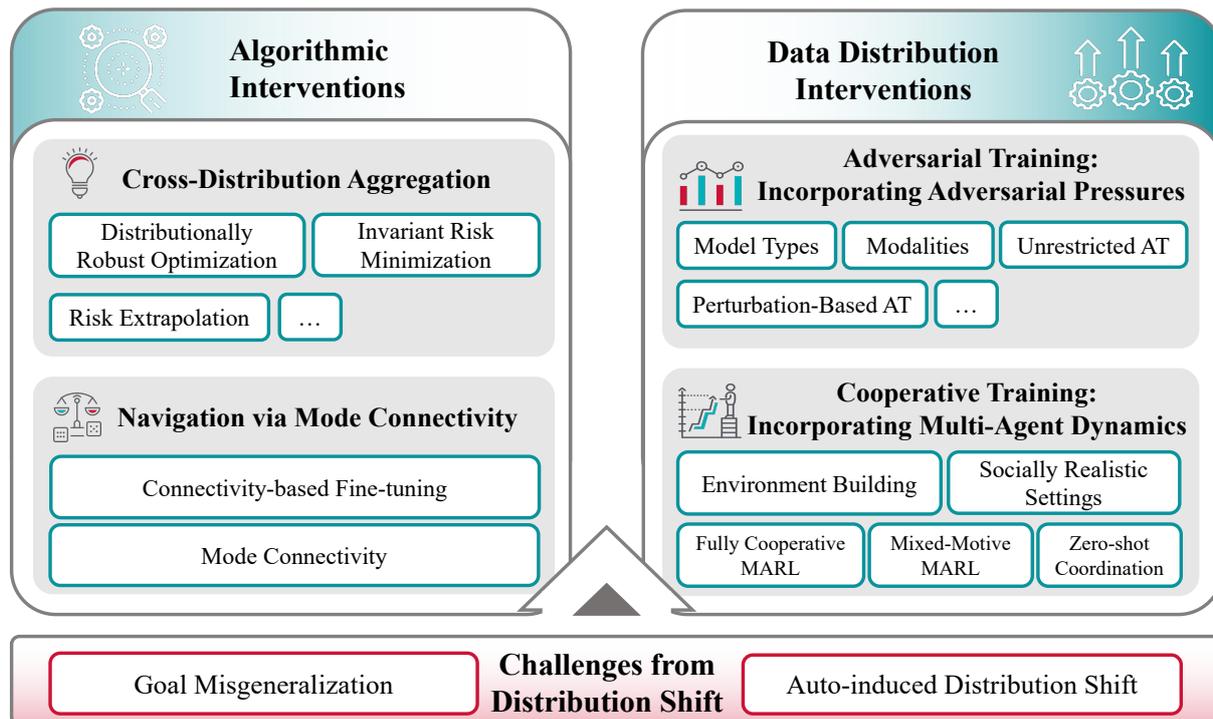
Figure 6: Framework of learning under distribution shift. The main challenges stemming from the distribution shift are goal misgeneralization and auto-induced distribution shift (§3.1). In our framework, we also introduce two kinds of methods to address distribution shift: algorithmic interventions (§3.2) that steer optimization during training, and data distribution interventions (§3.3) that expand the training distribution in a targeted manner by introducing real-world elements.

the generalization of *objectives* across distributions, as opposed to the generalization of *capabilities* (Di Langosco et al., 2022; Ngo et al., 2024).

We mainly discuss the preservation of alignment properties when learning under distribution shift in this section. We start the discussion by introducing the alignment challenges from distribution shift (§3.1). Subsequently, we delve into methods for addressing distribution shift, and discuss two approaches in particular: (1) algorithmic interventions (§3.2) that steer optimization during the training process, and (2) data distribution interventions (§3.3) that expand the training distribution by introducing specific elements into the training process, including adversarial training (Yoo and Qi, 2021; Bai et al., 2021; Ziegler et al., 2022) and cooperative training (Dafoe et al., 2021) (§3.3.2). Our framework for learning under distribution shift is shown in Figure 6.

3.1 The Distribution Shift Challenge

Before introducing the specific techniques, we initially demonstrate why one of the primary challenges in alignment is learning under distribution shift, and more specifically, the preservation of *alignment properties* (*i.e.*, adherence to human intentions and values) under distribution shift. We introduce two alignment challenges concerning the issue of distribution shift, namely goal misgeneralization (Di Langosco et al., 2022) and auto-induced distribution shift (ADS) (Krueger et al., 2020).

The training of AI systems optimizes for their adherence to the pursuit of the training reward/loss under the training input distribution. However, this adherence may not generalize to cases where the input distribution undergoes qualitative changes, *i.e.*, distribution shift. These changes include, for example, adversarial pressures (Poursaeed et al., 2021), multi-agent interactions (Critch and Krueger, 2020), and complicated tasks that human overseers cannot efficiently evaluate (Di Langosco et al., 2022), and reward mechanisms that can be gamed or manipulated (Krueger et al., 2020).

It's worth distinguishing two different failure modes here: goal misgeneralization (Di Langosco et al., 2022), in which the original and shifted distributions are given, and auto-induced distribution shift (Krueger et al., 2020), where the AI system alters the data distribution with its own behaviors in pursuit of reward.

Goal Misgeneralization This kind of challenge refers to the scenario where AI systems perform perfectly in the training distribution, but the capabilities learned in training distribution fail to generalize in OOD deployment,

and AI may present the pursuit of goals that are not in accordance with human wishes (Di Langosco et al., 2022). Goal misgeneralization[30] is to be distinguished from other forms of misgeneralization (*e.g.*, capability misgeneralization) where the agent becomes incompetent in OOD settings; instead, agents with goal misgeneralization *competently* pursue an *unwanted* goal in OOD settings.

A simplistic example is the case of *spurious correlations* (or *shortcut features*) (Geirhos et al., 2019; Di Langosco et al., 2022). For example, in an image classification dataset, green grass is a highly predictive feature for the label *cow*. However, it is essential to note that this feature needs to be more consistent and reliable across various data distributions (Murphy, 2023). Moreover, the causal confusion (*i.e.*, ignorant of the causal structure of the interaction between the advisor and the environment) in IL can result in goal misgeneralization (De Haan et al., 2019; Tien et al., 2022).

One major danger from goal misgeneralization lies in the indistinguishability between "optimizing for what human really wants" and "optimizing for human thumbs-ups";[31] the latter includes potentially deceiving or manipulating human evaluators (Shevlane et al., 2023) to receive their thumbs-ups. For example, Amodei et al. (2017) discovered that in a task where a robotic hand is supposed to grasp a small ball, the robotic hand fakes the action by using parallax in front of the lens to appear as if it has grasped the ball, without actually doing so. This behavior deceives the human annotator into thinking that the task has been completed.

When an AI system is trained or finetuned with human feedback, it is impossible to distinguish the two goals since both perform perfectly in training, and it is unclear which one the AI system will learn. In fact, during training, the human evaluators might be deceived or manipulated, implying that the AI system may be more strongly incentivized to optimize for human thumbs-ups rather than what the human wants. Current examples of this phenomenon exist in recommender systems (Kalimeris et al., 2021; Adomavicius et al., 2022), LLMs (Perez et al., 2023), and RL systems (Amodei et al., 2017).

Finally, one failure mode closely related to goal misgeneralization is the misalignment of *mesa-optimizers* (Hubinger et al., 2019c), where the ML model with learned model weights performs optimization within itself during inference ("mesa-optimization") (Hubinger et al., 2019c; Dai et al., 2023), and the objective of this optimization is not aligned with the model's training objective.

Auto-Induced Distribution Shift (ADS)   While training AI systems, we often consider the strengths and weaknesses of the agents themselves only and overlook the impact that these agents have on the environment. Past research often assumed that data is independently and identically distributed (Besbes et al., 2022), ignoring the effect of algorithms on data distribution. However, Krueger et al. (2020) posited that, in reality, agents could influence the environment during the decision-making and execution process, thus altering the distribution of the data generated by the environment. They referred to this type of issue as ADS. A real-world example is in recommendation systems, where the content selected by the recommendation algorithms might change users' preferences and behaviors, leading to a shift in user distribution. The distribution shift, in turn, further affects the output of the recommendation algorithms (Carroll et al., 2022). As AI systems increasingly impact the world, we also need to consider the potential further impacts on the data distribution of the entire society after agents are integrated into human society.

3.2   Algorithmic Interventions

When illustrating the algorithmic intervention methods, we first outline two classes of methods that steer optimization on various distributions during training to relieve distribution shift, namely, cross-distribution aggregation (§3.2.1) and navigation via mode connectivity (§3.2.2).

In the first part, we cover methods ranging from the initial approach of *empirical risk minimization* (ERM) (Vapnik, 1991) to *risk extrapolation* (REx) (Krueger et al., 2021), a method conceived to mitigate issues arising from models' dependence on spurious features. In the second part, we introduce *connectivity-based fine-tuning*, which guides the navigation of the loss landscape during training to encourage convergence upon non-spurious correlations, and which does so using insights from *mode connectivity* (Lubana et al., 2023).

3.2.1   Cross-Distribution Aggregation

One of the main reasons for distribution shift is spurious correlations in the model that are distinct from core objectives (Geirhos et al., 2019). By integrating learning information of different domains (or different distributions) into the optimization objective, we expect the model to learn truthful information and invariant relationships. In the following paragraphs, we first introduce ERM as the background and then introduce some methods to directly learn how to address distribution shift by integrating loss landscapes of different distributions in the training process.

---

[30]More examples of goal misgeneralization exist (DeepMind, 2020).

[31]Here, *human thumbs-ups* refer to high-reward feedback from human advisors or environment. However, AI systems may deliberately follow human preferences or deceive to get high rewards from humans, but actually don't really learn intended goals (*i.e.*, what human really wants).
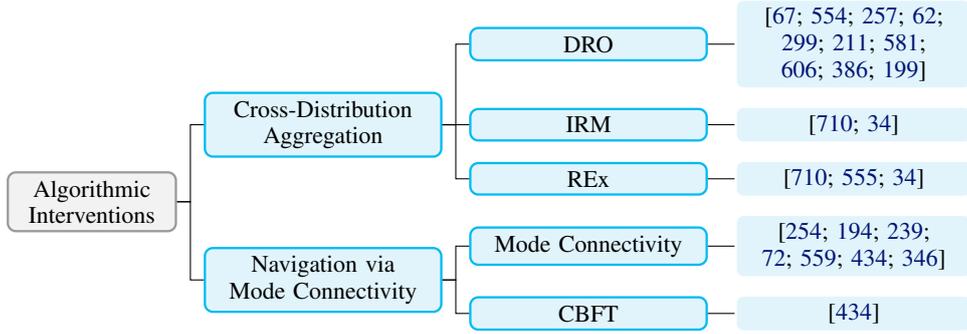
Figure 7: A tree diagram summarizing the key concepts and literature related to Algorithmic Interventions. The root node represents Algorithmic Interventions that aim to steer optimization during the training process. The main branches represent two main methods, namely cross-distribution aggregation (which aims to minimize risks on different distributions during training to find a predictor based on the invariant relationship instead of spurious features) and navigation via mode connectivity (which aims to fine-tune based on mode connectivity to enhance model generalization performance). Further sub-branches list vital techniques such as Distributionally Robust Optimization (DRO), Invariant Risk Minimization (IRM), Risk Extrapolation (REx), and Connectivity-based Fine-tuning (CBFT).

Empirical Risk Minimization (ERM)   Consider a scenario where a model has been developed to identify objects by their features effectively. The optimization target can be expressed as:

$$R(\boldsymbol{w}) = \int L\big(y, f(x, \boldsymbol{w})\big) d P(x, y)$$

where $L(y, f(x, \boldsymbol{w}))$ denotes the loss between data labels y and model outputs $f(x, \boldsymbol{w})$, while $P(x, y)$ signifies the target data distribution (Vapnik, 1991).

Nevertheless, a bias often exists between the dataset and the real world, implying that the features learned from the dataset may not necessarily be the ones we intend for the model to acquire. ERM is a strategy employed in statistical methods to optimize this bias. It operates on the assumption that, given the inaccessibility of the real-world target data distribution, the empirical data within the dataset should, ideally, closely approximate this unknown target distribution (Vapnik, 1991; Zhang et al., 2018b). In this context, the objective function is optimized and is redefined as:

$$E(\boldsymbol{w}) = \frac{1}{l} \sum_{i=1}^{l} L\big(y_i, f(x_i, \boldsymbol{w})\big)$$

where l can be different examples in one training distribution or different training distributions.

Minimizing the objective function above allows the model to learn the invariant relationship in different distributions. Naive ERM makes the naive assumption that the data is sampled from the target data distribution. However, if a significant discrepancy exists between the source distribution (or training distribution) and the target distribution, severe generalization issues can still arise (Szegedy et al., 2013).

Distributionally Robust Optimization (DRO)   Numerous studies posit that the sensitivity to distribution shift often arises from reliance on *spurious correlations* or *shortcut features* unrelated to the core concept (Geirhos et al., 2019; Hendrycks and Dieterich, 2018). For instance, models may judge based on background features rather than employing the correct features within the image (Geirhos et al., 2019; Beery et al., 2018). Building upon the foundations laid in prior research (Ben-Tal et al., 2009; Peters et al., 2015; Krueger et al., 2021), OOD Generalization can be formulated as follows:

$$r_{\mathcal{D}}^{\text{OOD}}(\boldsymbol{\theta}) = \max_{e \in \mathcal{D}} r_e(\boldsymbol{\theta})$$

This optimization seeks to enhance worst-case performance across a perturbation set, denoted as $\mathcal{D}$, by reducing the maximum value among the risk function set $\{r_e | e \in \mathcal{D}\}$. In *Distributionally Robustness Optimization (DRO)* (Duchi et al., 2021), the perturbation set covers the mixture of different domains' training distributions, and by minimizing the above objective function, we expect the model can find the invariant relationship between different training distributions. However, it should be noted that naively applying DRO to overparameterized neural networks may lead to suboptimal outcomes (Sagawa et al., 2020). Therefore, combining DRO with increased regularization

techniques such as l_2 penalty (Cortes et al., 2009) or early stopping (Prechelt, 2002) can substantially improve generalization performance. For more details on DRO, see *e.g.*, Rahimian and Mehrotra (2019); Sagawa et al. (2020); Lin et al. (2022a)

Invariant Risk Minimization (IRM)   Arjovsky et al. (2019) introduces an innovative learning paradigm to estimate nonlinear, invariant, causal predictors across diverse training environments, thereby facilitating robust OOD generalization. IRM aims to train a predictive model with solid performance across various environments while demonstrating reduced susceptibility to relying on spurious features. IRM can be considered an extension of Invariant Causal Prediction (ICP) (Peters et al., 2015), which involves hypothesis testing to identify the direct causal features that lead to outcomes within each specific environment instead of indirect features. IRM further extends ICP to scenarios characterized by high-dimensional input data, where variables may lack clear causal significance. The fundamental idea underlying IRM is that when confronted with many functions capable of achieving low empirical loss, selecting a function that exhibits strong performance across all environments is more likely to get a predictor based on causal features rather than spurious ones (Murphy, 2023).

Risk Extrapolation (REx)   The basic form of REx involves robust optimization over a perturbation set of extrapolated domains (MM-REx), with an additional penalty imposed on the variance of training risks (V-REx) (Krueger et al., 2021). By reducing training risks and increasing the similarity of training risks, REx forces the model to learn the invariant relationship in different domain distributions.

Amplifying the distributional variations between training domains can diminish risk changes, thereby enforcing the equality of risks. Taking CMNIST (Arjovsky et al., 2019) as an example, even though establishing a connection between color and labels is more straightforward than connecting logits and labels, increasing the diversity in color can disrupt this *spurious correlations* (or shortcut features) and aid the model in learning the genuine invariant relationship between logits and labels. Following previous research (Vapnik, 1991; Peters et al., 2017; Krueger et al., 2021), REx can be formulated as follows: Firstly, the Risk Function can be defined as follows:

$$r_e(\boldsymbol{\theta}) \doteq \mathbb{E}_{(x,y) \sim P_e(X,Y)} L\big(f_{\boldsymbol{\theta}}(x), y\big)$$

where $L(\cdot)$ represents a fixed loss function, and distinct training domains or environments can be formulated as the $P_e(X, Y)$ distribution. Next, the MM-REx term can be modeled as:

$$r_{\text{MM–REx}}(\boldsymbol{\theta}) = (1 - m\lambda_{\min}) \max_e r_e(\boldsymbol{\theta}) + \lambda_{\min} \sum_{e=1}^{n} r_e(\boldsymbol{\theta})$$

where n represents the number of distinct distributions or domains, and $\lambda_{\min}$ governs the extent of risk extrapolation. Moving on to the V-REx term, it can be modeled as:

$$r_{\text{V–REx}}(\boldsymbol{\theta}) = \alpha \operatorname{Var}\big(\{r_1(\boldsymbol{\theta}), \ldots, r_n(\boldsymbol{\theta})\}\big) + \sum_{e=1}^{n} r_e(\boldsymbol{\theta})$$

where $\alpha \geq 0$ controls the trade-off between risk reduction and enforcing risk equality.

In the MM-REx term, the $\lambda_{\min}$ can set nearly $-\infty$; therefore, the loss of specific domains may be high, meaning that the model may learn the spurious correlations. Minimizing the MM-REx and V-REx can reduce training risks and increase the similarity of training risks, encouraging the model to learn invariant relationships. Furthermore, REx has shown significant promise in experimental settings (Krueger et al., 2021), particularly in causal identification, making it a compelling approach for achieving robust generalization.

3.2.2   Navigation via Mode Connectivity

Following the above discussion about cross-distribution aggregation, in this section, we introduce mode connectivity as the prerequisite content. Then, we primarily discuss the Connectivity-Based Fine-Tuning (CBFT) (Lubana et al., 2023) method, illustrating how mode connectivity navigates the model to predict based on invariant relationships instead of spurious correlations by changing few parameters.

Mode Connectivity   Mode connectivity refers to the phenomenon where one can identify a straightforward path within the loss function space that connects two or more distinct local minima or patterns (Garipov et al., 2018; Draxler et al., 2018). In line with prior research (Benton et al., 2021; Pittorino et al., 2022; Lubana et al., 2023), a formal definition can be defined as follows:

The model's loss on a dataset $\mathcal{D}$ is represented as $\mathcal{L}(f(\mathcal{D}; \boldsymbol{\theta}))$, where $\boldsymbol{\theta}$ denotes the optimal parameters of the model, and $f(\mathcal{D}; \boldsymbol{\theta})$ signifies the model trained on dataset $\mathcal{D}$. We define $\boldsymbol{\theta}$ as a minimizer of the loss on this dataset if $\mathcal{L}(f(\mathcal{D}; \boldsymbol{\theta})) < \epsilon$, where ϵ is a small scalar value.

Minimizers θ_1 and θ_2, achieved through training on dataset $\mathcal{D}$, are considered to be mode-connected if there exists a continuous path γ from θ_1 to θ_2 such that, as θ_0 varies along this path γ, the following condition is consistently upheld:

$$\mathcal{L}\big(f(\mathcal{D};\theta_0)\big) \leq t \cdot \mathcal{L}\big(f(\mathcal{D};\theta_1)\big) + (1-t) \cdot \mathcal{L}\big(f(\mathcal{D};\theta_2)\big), \quad \forall t \in [0,1].$$

In essence, mode connectivity entails consistently finding a connecting pathway among minimizers in the parameter space, traversing regions of low loss without delving into regions of highly high loss. This implies that even when making minor adjustments to the model's parameters within the parameter space, the model's performance can remain relatively stable, mitigating significant performance degradation (Garipov et al., 2018). This concept lays the foundation for designing more effective optimization algorithms, enabling models to share knowledge and experiences across different tasks, enhancing both model performance and generalization capabilities.

Furthermore, we can define two models as mechanistically similar if they employ the same attributes of inputs for making predictions. Some research has demonstrated that the absence of linear connectivity implies mechanistic dissimilarity, suggesting that simple fine-tuning may not suffice to eliminate spurious attributes learned during the pre-training phase (Lubana et al., 2023; Juneja et al., 2022). However, it is promising to address non-linearly connected regions through fine-tuning, thereby effectively modifying the model's mechanisms to resolve the issue of OOD misgeneralization.

Connectivity-Based Fine-tuning (CBFT)   As discussed above, recent research has suggested that the absence of linear connectivity between two models implies a fundamental mechanistic dissimilarity. Lubana et al. (2023) finds that models tend to develop similar inference mechanisms when trained on similar data. This could be a significant reason for the emergence of bias in models, such as relying on the background information of images for classification rather than the objects depicted in the images. If this model mechanism is not adjusted during the finetuning process, the model may rely on these false attributes. To overcome this problem, they propose a valid strategy for altering a model's mechanism, which aims to minimize the following loss:

$$\mathcal{L}_{\mathrm{CBFT}} = \mathcal{L}_{\mathrm{CE}}\big(f(\mathcal{D}_{\mathrm{NC}};\theta), y\big) + \mathcal{L}_{\mathrm{B}} + \frac{1}{K}\mathcal{L}_{\mathrm{I}}$$

where the original training dataset is denoted as $\mathcal{D}$, and we assume that we can obtain a minimal dataset without spurious attribute C, denoted as $\mathcal{D}_{\mathrm{NC}}$.

Besides $\mathcal{L}_{\mathrm{CE}}$ that denotes the cross-entropy loss between model's prediction $f(\mathcal{D}_{\mathrm{NC}};\theta)$ and the ground truth label y, CBFT has two primary objectives: (1) The first objective entails modifying a model's underlying mechanism by repositioning it within the loss landscape, breaking any linear connection with the current minimizer. This is accomplished by maximizing $\mathcal{L}_{\mathrm{B}}$, referred to as the *barrier loss*. (2) The second objective involves mitigating reliance on spurious attributes in the original training dataset. This is achieved by optimizing $\mathcal{L}_I$, enabling the discovery of invariant relationships without the need for C. CBFT holds promise for shifting the mechanism from predicting objectives by spurious features to true features, just changing partial parameters of models.

3.3   Data Distribution Interventions

Besides algorithmic optimization, methods that expand the distribution of training data to include real-world elements can also reduce the discrepancy between training and deployment distributions. In this section, we specifically focus on the introduction of adversarial pressures and multi-agent dynamics.

3.3.1   Adversarial Training

AI systems can suffer from a lack of adversarial robustness, meaning that certain inputs designed to make them fail cause the models to perform poorly (Zheng et al., 2016), which has been shown in images (Huang et al., 2017) and texts (Zou et al., 2023b; Shah et al., 2023), as well as changes to semantic features in images (Geirhos et al., 2019; Bhattad et al., 2019; Shamsabadi et al., 2020; Casper et al., 2022) and texts (Jia and Liang, 2017), and even examples generated entirely from scratch (Song et al., 2018b; Ren et al., 2020; Ziegler et al., 2022; Chen et al., 2024). These failure modes are covered in the *red teaming* section (§4.1.3). It's worth noting that in addition to the robustness of AI model policies, the robustness of reward models that govern the training of advanced AI systems is also of importance, as the gradient descent optimization process could be seen as an adversary that may exploit loopholes in the reward model, a phenomenon named *reward model overoptimization* that has been experimentally demonstrated (Gao et al., 2023).

We consider adversarial robustness a case of distribution shift failure caused partly by a mismatch between AI systems' training distribution (where the training inputs are not adversarially constructed) and testing distribution (where the example can be adversarially constructed). The method of *adversarial training* (Yoo and Qi, 2021; Bai et al., 2021; Ziegler et al., 2022) mitigates this problem by introducing adversarial examples into training input

through a variety of ways (Bai et al., 2021), thus expanding the training distribution and closing the distribution discrepancy.

Adversarial training, which is similar to adversarial attacks, first started in the settings of image classification (Engstrom et al., 2019a), but later expanded to a wide range of settings. In addition to vision models, adversarial training algorithms have been proposed for language models (Wang et al., 2019a; Liu et al., 2020; Ziegler et al., 2022), vision-language models (Gan et al., 2020; Berg et al., 2022), *etc.* In terms of the model type, adversarial training has been applied to classification models (Bai et al., 2021), generative models (Ziegler et al., 2022), and RL agents (Pinto et al., 2017; Tan et al., 2020).

There are two major types of adversarial training: *perturbation-based* and *unrestricted*.

- **Perturbation-based Adversarial Training**. Mirroring *perturbation-based adversarial attack* (see §4.1.3), perturbation-based adversarial training introduces adversarially perturbed examples (*i.e.*, small changes to a normal data input which are designed to reduce model performance) into training (Goodfellow et al., 2014). Techniques in this vein (Bai et al., 2021) include the baseline approach of adding a regularization term into the loss function to assess model performance on a gradient-based perturbated input (Goodfellow et al., 2014), unsupervised (Carmon et al., 2019) or self-supervised (Hendrycks et al., 2019) approaches, and various supplemental techniques such as the introduction of curriculum learning which gradually intensifies adversarial pressure during training.

- **Unrestricted Adversarial Training**. Mirroring *unrestricted adversarial attack* (see §4.1.3), unrestricted adversarial training generalizes perturbation-based adversarial training to include *any* adversarial example that can fool the model, not necessarily ones obtained by adding a small amount of noise to another example. This includes *generative adversarial training*, which uses generative models to produce arbitrary adversarial inputs from scratch (Poursaeed et al., 2021), and the addition of syntactically or semantically modified adversarial examples to training input (Ziegler et al., 2022; Mao et al., 2022) which surprisingly eliminates the negative effects on the model's non-adversarial performance. Most works on unrestricted adversarial attacks also apply to unrestricted adversarial training (see §4.1.3 for an overview) and form an important part of the unrestricted adversarial training methodology.

3.3.2   Cooperative Training

Cooperative AI (Dafoe et al., 2020, 2021) aims to address uncooperative and collectively harmful behaviors from AI systems (see §1.1.2). The lack of cooperative capabilities in AI systems can be seen as a form of failure under distribution shift – systems are trained in single-agent settings that are qualitatively different from the real world, which could be massively multi-agent. This difference is indeed a difference in data distribution since the presence of other agents in the environment qualitatively alters the environmental state transition dynamics, leading to changes in the joint distribution of observations and rewards. We approach the problem by expanding our training distribution to include multi-agent interactions via *cooperative training*.

We introduce the branch of cooperative AI (what we call *cooperative training*) that focuses on specific forms of Multi-Agent Reinforcement Learning (MARL) training and complements formal game theory approaches in §4.3.1. The MARL branch of cooperative training tends to emphasize the AI system's *capabilities* for coordination (*e.g.*, coordination of a robot football team (Ma et al., 2022)), as opposed to *incentives* of cooperation (*e.g.*, mitigating failure modes like the prisoner's dilemma (Phelps and Russell, 2023)) which are the focus of the game theory branch. Here, we only cover the MARL branch due to its relevance to expanding training data distribution.

The field of MARL had traditionally been divided into the three branches of *fully cooperative* (where all agents share the same reward function), *fully competitive* (where the underlying rewards constitute a zero-sum game), and *mixed-motive* settings (where the reward incentives are neither fully cooperative nor fully competitive, corresponding to general-sum games) (Gronauer and Diepold, 2022). Among them, fully cooperative and mixed-motive settings are the most relevant for cooperative AI, and the latter has been especially emphasized due to its relative neglectedness (Dafoe et al., 2020). We also cover other research fronts, including zero-shot coordination (Hu et al., 2020; Treutlein et al., 2021), environment-building (Leibo et al., 2021), and socially realistic settings (Du, 2023).

- **Fully Cooperative MARL**. Fully cooperative settings of MARL are characterized by a shared reward function for all agents (Gronauer and Diepold, 2022). This unity allows us to completely disregard issues of cooperation *incentives* (since all incentives are perfectly aligned) and instead focus on effectively achieving the shared goal via coordination. Commonly adopted approaches (Oroojlooy and Hajinezhad, 2023) lie on a spectrum of centrality – from the baseline solution of purely independent training (Tan, 1993) to the approach of supplementing independent training with decentralized communications (Foerster et al., 2016), and then to *value factorization* which decomposes a global reward and determine each individual agent's contribution (Guestrin et al., 2001; Sunehag et al., 2018).
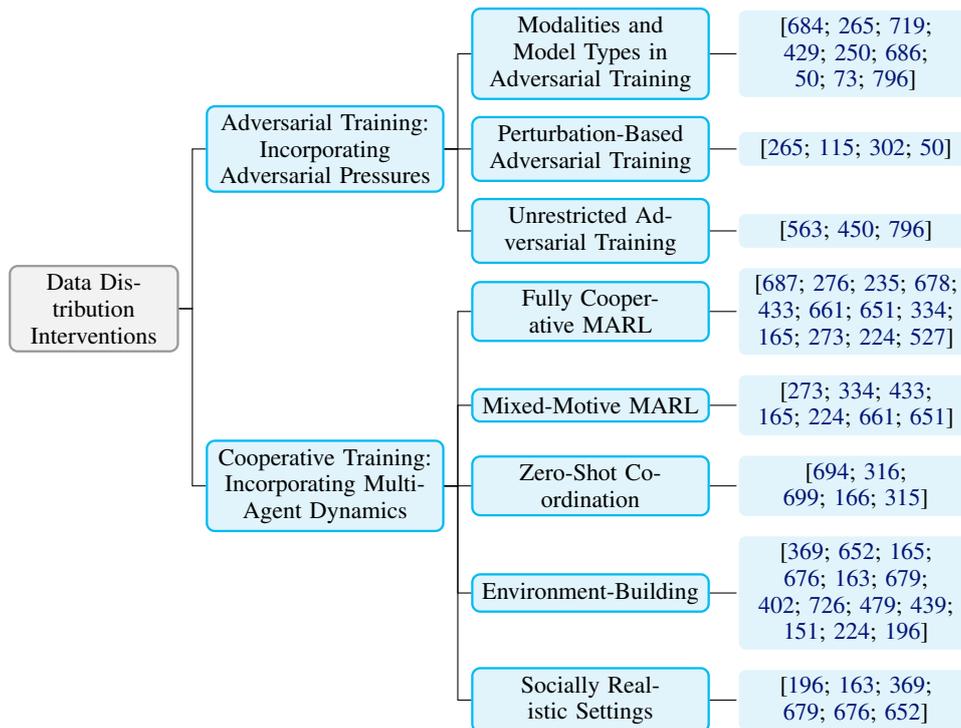
Figure 8: A tree diagram summarizing the key concepts and literature related to Data Distribution Interventions. The root node represents Data Distribution Interventions that try to combine multiple distributions during training, for example, adversarial examples and multi-agent interaction. The main branches represent promising methods, namely, Adversarial Training that incorporates adversarial pressures and Cooperative Training that incorporates multi-agent dynamics. Further sub-branches list key techniques such as perturbation-based and unrestricted adversarial training, and cooperative methods also include environment-building, socially realistic settings, zero-shot coordination, and other Multi-Agent Reinforcement Learning (MARL)-based techniques.

- **Mixed-Motive MARL**. Mixed-motive settings of MARL are characterized by a mixture of cooperative and competitive incentives – rewards for agents are not identical but aren't zero-sum either (Gronauer and Diepold, 2022). This includes game environments where teams play against each other (Jaderberg et al., 2019) and more nuanced settings such as negotiation (Cruz et al., 2019; FAIR et al., 2022). Examples of techniques for mixed-motive MARL, again ordered from decentralized to centralized, include using IRL-like methods to learn from human interactions (Song et al., 2018a), making communications strategic and selective (Singh et al., 2018) and adapting actor-critic methods by granting the critic access to global information (Lowe et al., 2017).

- **Zero-shot Coordination**. Zero-shot coordination is the goal of making AI systems able to coordinate effecively with other agents (including human agents) without requiring being trained together or otherwise being designed specifically to coordinate with those agents (Hu et al., 2020; Treutlein et al., 2021) – human beings who are complete strangers can still cooperate effectively, and we hope that AI systems can do the same. Early works were published under the name *ad hoc coordination*, covering evaluation (Stone et al., 2010), game-theoretic and statistical approaches (Albrecht and Ramamoorthy, 2013), and human modeling (Krafft et al., 2016). Recent advances include *other-play* (Hu et al., 2020) which randomizes certain aspects of training partners' policies to achieve robustness,[32] the introduction of multi-level recursive reasoning (Cui et al., 2021), and *off-belief learning* (Hu et al., 2021) which eliminates arbitrary conventions in self-play by interpreting partners' past actions as taken by a non-collusive policy.

- **Environment-building**. Game environments have been popular settings for cooperative training, including, for example, Hanabi (Muglich et al., 2022), Diplomacy (Cruz et al., 2019; FAIR et al., 2022), and football (Ma et al., 2022). On the more simplistic end, game theory models, especially those based on classical multi-agent dilemmas, have also been a popular choice of environment (Wang and Beliaev, 2021; Christoffersen et al., 2023). Also, Melting Pot (Leibo et al., 2021), a framework and suite of multi-agent environments, has been designed specifically for cooperative AI research. There has also been research on *unsupervised environment*

---

[32]This is in a similar spirit to *domain randomization* (Tobin et al., 2017).
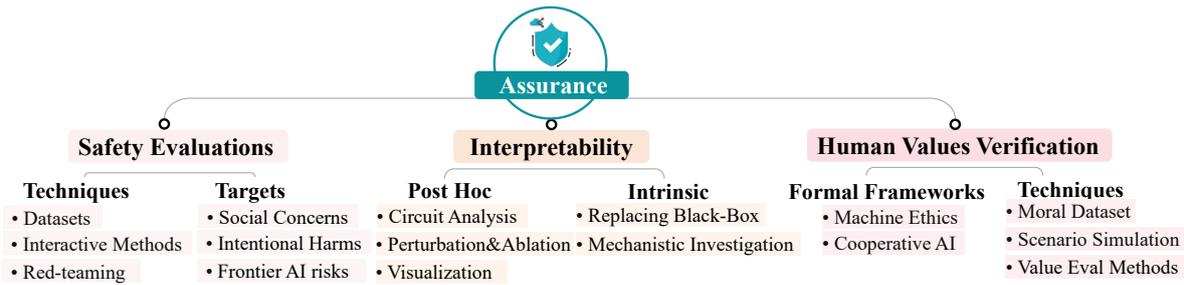
Figure 9: Our organization of research directions, techniques, and applications in assurance. We divide this section into *three* parts: Safety Evaluations – evaluation of AI systems' safety, which refers to the mitigation of accidents and harmful events caused by the AI system; Interpretability – making AI systems as well as its decision process more understandable to human beings; Human Value Verification – verifying whether AI systems can adhere to social and moral norms. The figure also displays the intricate logic of these sections.

design, which aims for a partial automation of the environment-building process (Dennis et al., 2020; Jiang et al., 2021).

- **Socially Realistic Settings**. It has been proposed that cooperative AI research should focus more on socially realistic environments (Du, 2023), which tend to be massively multi-agent (including both AI agents and human agents) and are highly diverse in both the composition of agents and modes of interactions. Implications of this vision (Critch and Krueger, 2020) include, but aren't limited to, building more realistic and open-ended environments (Klügl et al., 2005; Lehman et al., 2008; Wang et al., 2019b; Suo et al., 2021), scaling up MARL (Sun et al., 2020; Du, 2023), and incorporating new means of control such as social institutions and norms (Singh, 2014).

4 Assurance

Assurance refers to the measurement and refinement of AI systems' practical alignment after AI systems are actually trained or deployed (Batarseh et al., 2021). In this section, we categorize assurance into three parts based on a certain logic: Safety Evaluations – Evaluating AI systems on minimizing accidents during task execution as a basic need of assurance, Interpretability – Ensuring that humans can understand the decision-making process of AI systems and therefore assuring the safety and interoperability beyond evaluation, Human Value Verification – Verifying whether AI systems can align with human values, ethics, and social norms and satisfying the high-level need of AI systems' integration to the human society, as is described in the Figure 9.

In addition to methods that aim to *determine* if AI systems are safe and aligned, there are also assurance methods that actively *intervene* in the AI system or its deployment process to ensure such properties.

Machine Unlearning Datasets for model pretraining contain various types of undesirable and potentially dangerous content, including but not limited to information about bioweapons and cyberattack (Hendrycks et al., 2021b). The field of *machine unlearning* has aimed to remove such knowledge after a model is trained (Bourtoule et al., 2021). Compared to direct filtering of the training dataset, this approach faces more technical challenges, but it retains more flexibility in deployment and also allows categorical removal of a given piece of information (Eldan and Russinovich, 2023). Dataset filtering and unlearning ought to be seens as complementary approaches that work best together.

Controlling Unaligned Systems While complete alignment may be difficult, it is still possible to safely utilize unaligned models if their extent of misalignment is limited and if we have access to supervisor AI systems. Algorithmic procedures have been developed to minimize probabilities of failure when given trusted and untrusted systems with differing capabilities (Greenblatt et al., 2023). In general, alignment-focused *process engineering* of deployment procedures could be a valuable direction to explore.

We then go on two review the three categories of alignment measurement efforts.

4.1 Safety Evaluations

Safety refers to mitigating accidents caused by design flaws in AI systems and preventing harmful events that deviate from the intended design purpose of the AI system (Amodei et al., 2016). In fact, safety stands as a shared requirement across all engineering domains (Verma et al., 2010). Moreover, it holds particular importance in constructing AI systems, because of the characteristics of AI systems (Steinhardt, 2015). We categorize the safety of
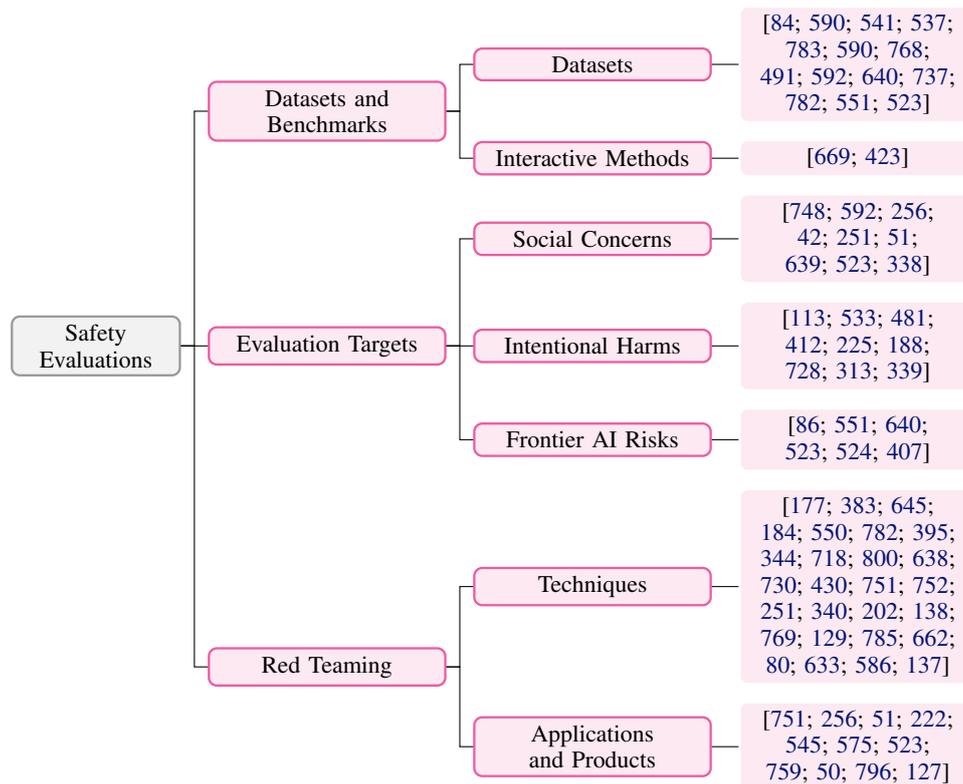
Figure 10: A Tree diagram summarizing the key concepts, logic, and literature related to Safety Evaluation. The root of the tree represents Safety Evaluation, which aims to *measure the accidents caused by design flaws in AI systems and harmful events that deviate from the intended design purpose of the AI system*. The main branches represent the main structure of safety evaluation, including Datasets and Benchmarks, Evaluation Targets, and Red Teaming techniques. Further sub-branches list key works exploring each of these branches. This diagram provides an overview of research directions and specific techniques for measuring AI systems' safety alignment degree.

AI systems into the following categories: *Social Concerns* refer to explicit and comparatively identifiable characteristics of safe AI systems, including aspects such as toxicity (Stahl and Leach, 2023), and *Intentional Behaviors* share the characterization of relatively complicated investigation and substantial potential harm, represented by power-seeking, deception, and other frontier AI risks (Shevlane et al., 2023).

Following the logic above, we start with the techniques to form datasets and benchmarks of safety evaluation in §4.1.1 and further explore the evaluation targets and their characteristics in §4.1.2. At the end of this section, we include the red-teaming technique §4.1.3, which assesses the AI system's robustness beyond evaluation.

4.1.1   Datasets and Benchmarks

In the discussions on safety evaluation, it is crucial to prioritize datasets and benchmarks as the cornerstone elements, so we first introduce the basic techniques to build datasets and benchmarks and then move on to newer interactive methods.

Dataset   Among all the assurance techniques, the dataset method could be considered the most elementary and straightforward one (Celikyilmaz et al., 2020). This method assesses the response of AI systems by presenting them with predefined contexts and tasks (Paullada et al., 2021), balancing the cost, quality, and quantity of data. Research on the dataset method encompasses data sources, annotation approaches, and evaluation metrics. Given that evaluation metrics can vary based on its subject (Sai et al., 2022), this section primarily emphasizes dataset sources and annotation methods.

- **Expert design**. In the early stage of a domain, expert design is widely used in building datasets, where experts create samples based on actual needs to ensure the dataset covers a wide range of potentially dangerous situations to form datasets (Roh et al., 2019). For instance, initial-stage datasets, *e.g.*, WEAT (Bolukbasi et al., 2016) and BBQ (Parrish et al., 2022) for bias detection used expert design to harvest a wide coverage and high accuracy while sharing the limitations in terms of cost and breadth, leading to the later development of more efficient methods.

40

- **Internet collection**. Previous expert design methods have the flaw of rather high cost and lower efficiency, and internet collection can obtain datasets that contain actual user-generated textual content on a rather large scale (therefore convenient for both training and testing), reflecting real-world text generation scenarios (Yuen et al., 2011), but the raw data collected also needs careful selection and annotation (Roh et al., 2019). Well-known instances of these datasets include OLID (Zampieri et al., 2019) and SOLID (Rosenthal et al., 2021) gathering original Twitter texts for toxicity assessment, WinoBias (Zhao et al., 2018) and CrowS-Pairs (Nangia et al., 2020) gather content potentially containing bias from the internet for further annotation. However, it's important to acknowledge that, as is also mentioned in Papernot et al. (2016), internet-collected datasets naturally carry risks such as privacy and safety concerns, so additional processing is necessary.

- **AI Generation**. The concept of autonomously generating datasets was explored relatively early, even before the emergence of elementary forms of LLMs (Weston et al., 2015). However, during this early stage, AI-generated datasets were limited by the capabilities of AI systems, so their quality was not as good as internet-collected and manually annotated datasets. It wasn't until LLMs reached relatively high levels of proficiency in logical reasoning context understanding and approached or surpassed human-level performance (OpenAI, 2023a) that LMs gained the ability to mimic the structure and logic of existing datasets to compose new ones. As is shown in papers such as Zhang et al. (2022) and Perez et al. (2023), AI systems have made progress in generating datasets for evaluation purposes, surpassing the quality of some classical datasets. However, according to these papers, this approach still faces limitations rooted in the capabilities of large models themselves, including issues like instruction misunderstanding and example diversity, which require further refinement.

Interactive Methods   Due to the static nature of datasets, they possess relatively fixed evaluation content and can be vulnerable to targeted training (Holtzman et al., 2019). Additionally, the evaluation content may not fully reflect the strengths and weaknesses of corresponding capabilities (Engstrom et al., 2020). As the demands for language model evaluation continue to escalate, new interactive assurance methods have emerged, which can be categorized into two groups: Agent as Supervisor and Environment Interaction.

- **Agent as Supervisor**. It is an assurance method that involves using an agent to assess the outputs of AI models. This evaluation approach is characterized by its dynamism and flexibility. Typically, there is a predefined framework for interaction between the agent and the AI system under evaluation (Cabrera et al., 2023). In this method, the agent can be a human participant engaged in experiments through an online system (Stiennon et al., 2020), a more advanced language model evaluating relatively less capable language models through multi-turn interactions (Lin and Chen, 2023), or in the context of *Scalable Oversight*, a less powerful but more trustworthy model (Greenblatt et al., 2023). This evaluation form offers advantages such as automation and lower cost compared to human agents.

- **Environment Interaction**. It aims to create a relatively realistic environment using elements such as humans and other LLMs to assess the alignment quality of AI models through multiple rounds of interaction (Liu et al., 2024b). One method is using peer discussions, where multiple LLMs engage in dialogue, to enhance evaluations of AI systems, particularly when their capabilities are relatively close to each other. Moreover, by building a world model (Li et al., 2022b), the generalization and exploration abilities of AI systems can be comprehensively evaluated.

4.1.2   Evaluation Targets

To achieve the goal of safety alignment, the assurance of AI systems can be divided into different small targets (Shevlane et al., 2023). The subsequent section gives an introduction to these subjects and, furthermore, discusses some of the domain-specific analyses of assurance methods within these realms, while the table 3 will show examples of alignment assurance works in these domains.

Toxicity   It refers to content in the output of AI systems that is unhelpful or harmful to humans (Sheth et al., 2022). Before the advent of advanced language models, early toxicity evaluation primarily focused on detecting toxic language and identifying harmful statements in an internet context, like the WCC (Wulczyn et al., 2017), which collected and manually labeled comments from Wikipedia discussion pages. With the emergence of pre-trained language models, assurance against toxicity adopted a prompt-generation paradigm to assess the risk of language models generating toxic content in response to specific prompts (Gehman et al., 2020; Ganguli et al., 2022; OpenAI, 2023a). However, in crowdsourced environments, annotation scores may vary by person, so relative labeling, where crowdsource workers select from two different answers during a chat, is needed to enhance crowdsource quality (Bai et al., 2022a). Furthermore, subsequent datasets (Ganguli et al., 2022; Ji et al., 2024b) employ a red teaming design pattern that induces toxic responses through adversarial inputs, further strengthening the assurance of model robustness.

Table 3: A Chart of Safety Evaluation Examples: Specific dataset works are listed in this chart, along with their detailed information: *evaluation targets*, *first release time*, *most recent update time* (we list them separately because some datasets are consistently being updated), *information quantity* (the sum of the information form unit), *institution*, *information form*, *baseline model* and *information source*. Moreover, to contain more information, we made some abbreviations in the chart: We shortened the release time and recent updates by concatenating the last two digits of the year and the month and only taking the institution of the paper's first author, and we use combinations of uppercase letters to replace long words in information form: SP for Sentence Pairs, SL for Sentence-Label, ST for sentence template, PP for pronoun pairs, and SS for single selections.

	Dataset	Release Time	Recent Update	Info Quantity	Institution	Information Form	Baseline Model	Infomation Source
Bias	Aequitas [608]	18/05	23/04	-	U.Chicago	Python	-	Self Build
	WinoS [598]	18/10	19/01	0.72K	JHU	ST	Rule&Neural	Self Build
	EEC [366]	18/05	-	8K	NRC Canada	SP	SVM	Selection
	GAP [729]	18/05	-	8.9K	Google	PP	Transformer	Wikipedia
	OLID [768]	19/05	-	14K	U.Wolver.	SL	SVM&LSTM	Twitter
	CrowS-Pairs [491]	20/03	21/10	1.5K	NYU	SP	BERT	MTurk
	StereoSet [486]	20/04	22/04	17K	MIT	SS	BERT&GPT-2	MTurk
	BBQ [541]	21/05	22/07	58.5K	NYU	SS	Multiple LLMs	MTurk
	LM-Bias [416]	21/07	22/01	16K	CMU	QA Pair	GPT-2	Corpus Select
	VQA-CE [173]	21/03	21/10	63K	Sorbonne	Multimodal	-	Self-Build
	AuAI [392]	23/01	-	-	Sorbonne	Framework	-	Self Build
Toxicity	WCC [748]	16/01	-	63M	Wikimedia	SL	Human	Wikipedia
	RTP [256]	19/10	21/04	100K	UW	Prompt	GPT-2	Refinement
	SOLID [592]	20/05	-	9M	IBM	SL	BERT	Twitter
	Toxigen [288]	20/05	23/06	274K	MIT	SL	GPT-3	GPT Gen.
	HH-RLHF [51]	22/04	22/09	162K	Anthropic	SP	Claude	Corpus Refine
	BeaverTails [338]	23/06	23/07	30K	PKU	QA Pair	Multiple LLMs	Corpus Refine
Power Seeking	MACHIAVELLI [533]	23/04	23/06	134	UCB	Games	GPT-4&RL	Selection
	BeaverTails [338]	23/06	23/07	30K	PKU	QA Pair	Multiple LLMs	Corpus Refine
Situation Awareness	SA Framework [611]	20/07	-	-	MIT	Framework	-	Self Build
	EWR [412]	-	-	10	Havard	Game	Othello GPT	Self Build
Hallucination	PARENT [188]	19/06	-	-	CMU	Metric	-	Self Build
	PARENT-T [728]	20/05	-	-	NYU	Metric	-	Self Build
	ChatGPT-Eval [58]	23/02	23/03	-	HKUST	Multimodal	ChatGPT	Integration
	POPE [415]	23/05	23/08	2K	RUC	Multimodal	Multiple LVLMs	Dataset Refine

Power-seeking   It is a kind of risk that AI systems may seek power over humans once they possess certain levels of intelligence (Turner et al., 2021). In Carlsmith (2022), the authors point out that AI systems already have the conditions for power-seeking, including advanced capabilities, agentic planning, and strategic awareness. However, the assurance against power-seeking is still in its early stages. One representative work in this area is the Machiavelli (Pan et al., 2023a), which constructs a benchmark consisting of decision-making games to assess whether AI systems can balance competition with moral ethics during the game. The conclusion of this work suggests that AI systems still struggle to balance achieving rewards with behaving morally, thus further research in this field is needed.

Deceptive Alignment   When the AI system is situationally aware, it may recognize that getting high rewards can preserve themselves by preventing significant gradient descent, therefore preserving its original goal (Hubinger et al., 2019c; Kenton et al., 2021; Ngo et al., 2024). This process is called *Deceptive Alignment*. In the current context, deceptive alignment is already achievable, as is proved by (Hubinger et al., 2024). Directly evaluating the deceptive alignment is difficult, for the pronoun *deceptive alignment* is naturally against the traditional train-evaluation loop. Thus, deceptive alignment might be discovered by indirect methods such as interpreting model parameters (see 4.2), or representation engineering (Zou et al., 2023a).

Moreover, deceptive alignment is closely related to *situational awareness*, *i.e.*, AI systems with a certain degree of prediction and understanding of the states and developments of entities in their working environment to make corresponding decisions. In (Li et al., 2022b), the authors evaluate the performance of language models in the board game Othello, showing that language models have the ability to predict possible future states within the action space in a nonlinear representation.

Hallucination   AI systems may generate information or responses that are not grounded in factual knowledge or data, leading to the creation of misleading or false content, which is formally called Hallucination (Ji et al., 2023). Hallucination evaluation aims to assure the consistency of the knowledge in the AI system's output with

the knowledge given by its training data and knowledge base (Ji et al., 2023; Zhang et al., 2023c). The earliest statistical-based hallucination evaluation methods used n-grams to directly calculate the overlap of vocabulary between the input and output content (Dhingra et al., 2019; Wang et al., 2020). However, this type of evaluation has a limitation: It only considers lexical overlap and does not take into account semantics or sentence meaning (Ji et al., 2023), making it unsuitable for evaluating more complex forms of hallucination. Later assurance methods shifted from statistical approaches to model-based methods, which are more robust compared to statistical token-difference-based methods (Honovich et al., 2021). While this evaluation method is more advanced than previous ones, it still has the limitation that the model can only output the degree of hallucination and may have difficulty pinpointing specific errors (Falke et al., 2019).

Frontier AI Risks   In addition to the assurance content described above, the enhancement of AI systems in recent years has given rise to a series of new assurance needs (OpenAI, 2023a). Currently, there is not much public information available for research on these assurance needs, hence this section will provide a brief introduction to some of the more significant ones:

- **Cyber Security & Biological Weapons**. Advanced LLMs may be misused for cyber-attacks, the production of bio-weapons, and other extremely harmful behaviors (Shevlane et al., 2023). Although GPT-4 cannot play a significant role in exploiting network vulnerabilities due to its limited context window, it has been proven to demonstrate strong capabilities in identifying network vulnerabilities and in social engineering (OpenAI, 2023a). Similarly, Lentzos (2022) have stated the robust abilities of AI systems in the field of bio-weapons and the military, highlighting the risks of misuse of such capabilities. It emphasizes the necessity to ensure that these models can identify and reject malicious requests.

- **Deception & Manipulation**. AI systems have the potential to negatively influence users by outputting text, including disseminating false information, syncopating humans, and shaping people's beliefs and political impacts (Shevlane et al., 2023; Sharma et al., 2024). Distinguished from hallucination, the misinformation here is not a flaw of the model itself but rather a deliberate action. Special assurance measures need to be designed for controlling these kinds of behavior.

- **Jailbreak**. It refers to the bypassing of AI systems' safeguard mechanisms by users, for example, by constructing specific types of input. This behavior can be limited to text (OpenAI, 2023a; Deng et al., 2023; Huang et al., 2024; Yong et al., 2023),[33] or it may take multi-modal forms (OpenAI, 2023b). Specifically, multi-modal jailbreaks make traditional text-based heuristic methods for identifying attack content infeasible, necessitating special multi-modal handling methods. Further discussion of jailbreak can be found in §4.1.3.

- **Self-Preservation & Proliferation**. This refers to the tendency of AI systems for self-protection and replication, and in this process, breaking the limit from their environment. These tendencies are examples of *instrumental sub-goals* (Bostrom, 2012). While this tendency can be beneficially harnessed, it is dangerous in the absence of regulation (Perez et al., 2023). This tendency has been emphasized and evaluated by various sources (Perez et al., 2023; Kinniment et al., 2023; OpenAI, 2023a,b).[33]

4.1.3   Red Teaming

Red teaming is the act of generating scenarios where AI systems are induced to give unaligned outputs or actions (*e.g.*, dangerous behaviors such as deception or power-seeking, and other problems such as toxic or biased outputs) and testing the systems in these scenarios. The aim is to assess the robustness of a system's alignment by applying adversarial pressures, *i.e.* specifically trying to make the system fail. In general, state-of-the-art systems – including language models and vision models – do not pass this test (Perez et al., 2022; Zou et al., 2023b; Liu et al., 2023; Chen et al., 2024).

In game theory and other fields, red teaming was introduced much earlier, and within computer science, the concept of red teaming was proposed in the security field, where it had a similar meaning of adversarially assessing the reliability and robustness of the system. Later, Ganguli et al. (2022); Perez et al. (2022) introduced this idea to the field of AI, and more specifically, alignment.

The motivation for red teaming is two-fold: (1) to gain assurance on the trained system's alignment, and (2) to provide a source of adversarial input for adversarial training (Yoo and Qi, 2021; Bai et al., 2021; Ziegler et al., 2022), probing models (Kalin et al., 2020), and further utilities. Here, we focus on the first. It's worth noting that the two objectives aren't separable; works targeting the first motivation also help provide a basis for the second.

---

[33]Relevant discussions in OpenAI (2023a) can be found in its *system card* appendix.

Reinforced, Optimized, Guided, or Reverse Context Generation   This category includes using various methods to generate coherent contexts (prompts) that are inducive to unaligned completions from the language model. Perez et al. (2022); Deng et al. (2022); Casper et al. (2023c) train or tune a separate language model with RL to make it generate desired prompts, which are then fed to the red-teamed model. Perez et al. (2022); Si et al. (2022) also uses other methods such as zero-shot, few-shot, or supervised finetuning-based generation. Lee et al. (2022); Jones et al. (2023) generates misalignment-inducive contexts by performing optimization on the prompt – bayesian optimization and discrete optimization, respectively. Dathathri et al. (2019); Krause et al. (2021) propose the method of guiding an LLM's generation using a smaller classifier; this is proposed in detoxification but is transferable to the red teaming context. Lastly, Zhang et al. (2022) generates misalignment-inducive contexts through *reverse generation, i.e. constructing adversarial contexts conditioned on a given response*, which can be seen as an inverse process for model inference.

Manual and Automatic Jailbreaking   As is defined above 4.1.2, *Jailbreaking* (Shen et al., 2023) is an informal term that refers to the act of bypassing a product's constraints on users – and in the case of LLMs, bypassing LLMs' tendencies to not answer misalignment-inducive questions, a feat of alignment training. Most existing attempts are scattered across the Internet in the form of informal reports and involve adding prefixes and suffixes to the original text (Zou et al., 2023b). Research has descriptively analyzed the existing attempts (Liu et al., 2023; Shen et al., 2023; Deng et al., 2023; Huang et al., 2024), as well as providing causal explanations for the phenomenon (Wei et al., 2024). In addition, past (Wallace et al., 2019) and current (Zou et al., 2023b; Shah et al., 2023) works have proposed effective methods to automatically generate such prompts, prefixes, or suffixes that nullify LLMs' tendencies to avoid misalignment-inducive questions.

Crowdsourced Adversarial Inputs   Several works (Xu et al., 2020, 2021; Ganguli et al., 2022) have produced misalignment-inducive prompts by crowdsourcing, *i.e.* recruiting human red teamers (possibly via online platforms) and instruct them to provide adversarial prompts. Besides, companies in the AI industry also build mechanisms to collect adversarial inputs, *i.e.* the red teaming network of OpenAI[34] and the bug hunter program of Google[35]. These methods (arguably) provide more flexibility and resemblance to real-world use cases but have higher costs and lower scalability.

Perturbation-Based Adversarial Attack   In the field of computer vision, there have been many works studying adversarial attacks on vision models that rest on the method of *perturbation, i.e.*, performing small perturbations to the pixel contexts of the image (usually bounded by a pixel-wise matrix norm) to make the model confidently produce false outputs on the perturbated image (Chakraborty et al., 2021). This type of adversarial attack has also been extended to language models (Jia and Liang, 2017; Ebrahimi et al., 2018; Zang et al., 2020; Cheng et al., 2020) and vision-language models (Zhao et al., 2024).

Unrestricted Adversarial Attack   *Unrestricted adversarial attack*, proposed in (Song et al., 2018b), is a more general form of adversarial attack. It removes all restrictions on the adversarial examples, and therefore, for instance, the adversarial example can be generated from scratch, as opposed to being generated from an existing example, as in the case of perturbation-based methods. Many methods for unrestricted adversarial attack have been proposed; the most notable ones include (Song et al., 2018b; Chen et al., 2024) which generate realistic adversarial images using generative models, and (Bhattad et al., 2019; Shamsabadi et al., 2020) which manipulates semantically meaningful traits such as color and texture. Unrestricted adversarial attack has also been extended to text classification models (Ren et al., 2020).

Datasets for Red Teaming   A number of works on red teaming and related topics have compiled datasets consisting of red teaming prompts or dialogues, including the IMAGENET-A and IMAGENET-O dataset (Hendrycks et al., 2021c), the BAD dataset (Xu et al., 2020), the red teaming section of HH-RLHF dataset (Bai et al., 2022a), and the Real Toxicity Prompts dataset (Gehman et al., 2020).

Existing Red Teaming Practices in Industry   The practice of red teaming is gaining popularity in the AI industry. Cases of adoption include OpenAI (who performed red teaming on its system GPT-4 to produce part of its System Card) (OpenAI, 2023a), NVIDIA (Pearce and Lucas, 2023), Google (Fabian, 2023), and Microsoft (Ram Shankar Siva Kumar, 2023). During an event at the DEF CON 31 conference, models from 9 companies undergo red teaming from the conference participants;[36] this red teaming event is held in partnership with four institutions from the U.S. public sector, including the White House.

---

[34] https://openai.com/blog/red-teaming-network
[35] https://bughunters.google.com/about/rules/6625378258649088
[36] https://www.airedteam.org/

Downstream Applications   Red teaming plays a crucial role in the adversarial training of AI systems by providing adversarial input (Yoo and Qi, 2021; Bai et al., 2021; Ziegler et al., 2022). In addition, adversarial examples produced from red teaming can also be used to interpret models (Casper et al., 2022).

4.2   Interpretability

Interpretability is a research field that makes machine learning systems and their decision-making process understandable to human beings (Doshi-Velez and Kim, 2017; Zhang and Zhu, 2018; Miller, 2019). Interpretability research builds a toolbox with which something novel about the models can be better described or predicted. In this paper, we focus on research that is most relevant to alignment and safety,[37] and empirically, those techniques make neural networks safer by studying the internal structures and representations of the neural networks (Räuker et al., 2023). Interpretability is an important research direction because in principle gaining safety guarantees about white-box systems is easier than black-box ones. The taxonomy of interpretability tools varies according to sub-fields and purposes (Doshi-Velez and Kim, 2017; Rudin, 2019). There are several ways to break down interpretability research:

- *Explainability and Transparency*. Explainability research aims to understand why models generate specific output, whereas transparency aims to understand model internals (Critch and Krueger, 2020).

- *Weights, Neurons, Sub-networks or Representations*. This classification organizes interpretability methods by seeing which part of the computational graph that method helps to explain: weights, neurons, sub-networks, or latent representations (Räuker et al., 2023).

- *Safety or the Science of Deep Learning*. Researchers also conduct interpretability research with different purposes: some do it to safely deploy AI systems, while others aim for a complete science of neural network. But the line gets blurred as mechanistic interpretability research aims for both (Olah et al., 2020; Olah, 2023).

- *Intrinsic and Post Hoc Interpretability*. By the stage of intervention, interpretability research is divided into intrinsic interpretability and post hoc interpretability (Carvalho et al., 2019): the former focuses on making intrinsically interpretable models, while the latter designs post hoc interpretability methods that offer explanations to model behaviors.

- *Mechanistic Interpretability, Representation Engineering, and Concept-based Interpretability*. Three research agendas have gained traction in the AI safety and alignment community (Impact, 2023): **Mechanistic Interpretability**, which, taking a bottom-up approach, aims to gain an understanding of low-level mechanics for algorithms implemented by neural networks (Olah et al., 2020), **Representation Engineering**, which, in contrast, taking a top-down approach, monitors (and manipulates) high-level cognitive phenomenon in neural networks (Zou et al., 2023a), and **Concept-based Interpretability** that locates learned knowledge representations in the neural networks, in contrast to what the models output (Meng et al., 2022a,b). The commonality among all three is linking a feature with a set of neurons simultaneously.

In this section, we adopt the *Intrinsic and Post Hoc Interpretability* classification method, for it offers a more generic framework suitable for various AI systems beyond neural network, and it divides the interpretability analysis both during the system designing and after the system has been deployed (Räuker et al., 2023), compared to other classification methods. Specifically, we discussed mechanistic interpretability techniques that take place in model designing and post hoc stages separately in post hoc and intrinsic interpretability subsections.

4.2.1   Intrinsic Interpretability

Researchers make deep learning models intrinsically more understandable, which is usually called *intrinsic interpretability* (Carvalho et al., 2019). In contrast to the symbolic approach, which emphasizes the creation of interpretable models, the modern deep learning approach tends to yield models with enhanced capabilities but potentially reduced interpretability. Compared to interpreting the black-box models, designing models that are intrinsically interpretable is safer and more efficient (Rudin, 2019). To make intrinsically interpretable models, the research community designs modular architecture, which is robust to adversarial attacks and free of superposition (Anthropic, 2022; Räuker et al., 2023). Notably, mechanistic interpretability, often regarded as a set of *post hoc interpretability techniques*, arguably facilitates the process of making more interpretable models.

---

[37]For a more comprehensive review of interpretability and its methods, we recommend (Räuker et al., 2023).
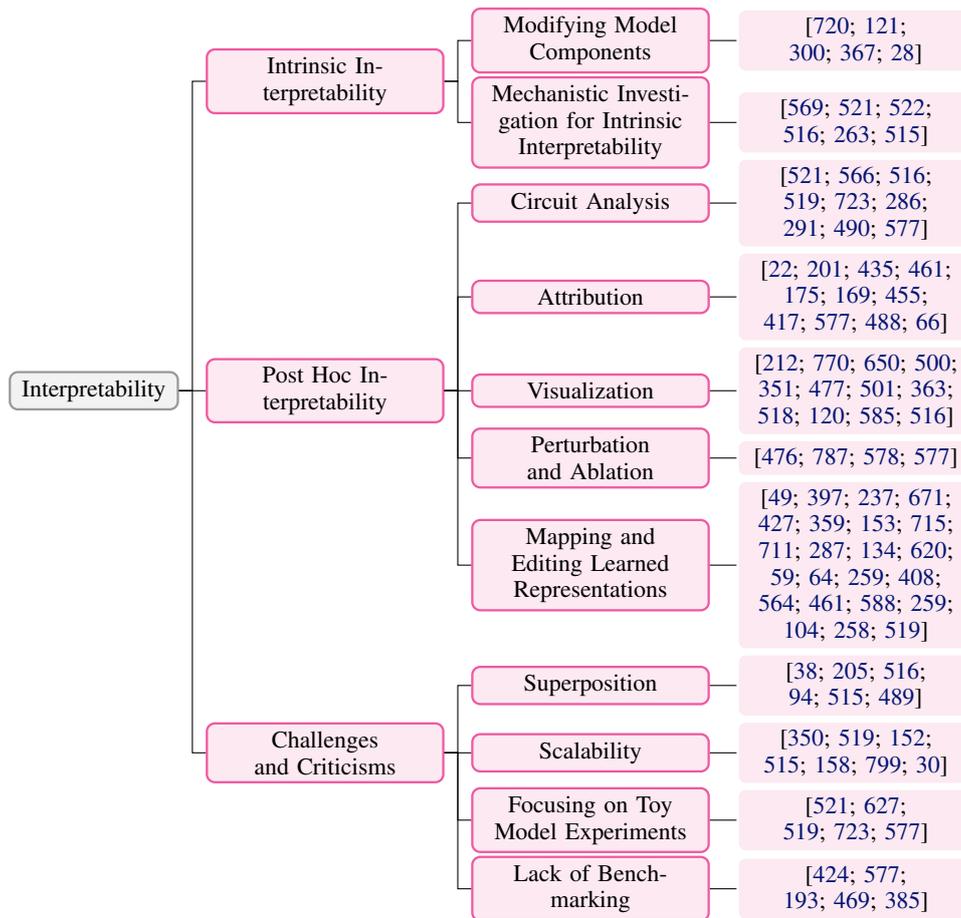
Figure 11: A Tree diagram summarizing the key techniques concepts, challenges, and literature related to Interpretability.

Modifying Model Components Model components, such as feedforward layers, are hard to interpret (*i.e.*, it's hard to articulate their behavior in human-understandable terms) because those layers have many polysemantic neurons that respond to unrelated inputs (Du et al., 2019). Thus, there are certain modifications applied to these back-box components and their related structures to make reverse engineering easier, and thus improve their interpretability (Carvalho et al., 2019). There are a number of existing works to encourage interpretable results by modifying loss functions (Ross et al., 2017), adding a special interpretable filter or embedding space (Zhang et al., 2018c; Wang et al., 2021), using dynamic weight depending on the input (Foerster et al., 2017), and modifying intermediate layers (Li et al., 2022a). Specifically, Lage et al. (2018) proposed a human-in-the-loop algorithm that directly utilizes human feedback to quantify the subjective concept, thus achieving more reliable results. In transformer models, Anthropic proposes SoLU to replace the activation function, increasing the number of interpretable neurons and making reverse engineering easier while preserving performance (Anthropic, 2022). This is still an early exploration as a potentially important line of work, and challenges remain, such as the scalability of this method (Anthropic, 2022).

Reengineering Model Architecture Modifying existing model components is beneficial to reverse engineering (Carvalho et al., 2019; Foerster et al., 2017), but they cannot make models *fully understandable*, so some researchers started to reengineer model architecture to build theoretically interpretable models (Carvalho et al., 2019; Mascharka et al., 2018). Notably, it is generally believed that there exists a trade-off between model interpretability and its performance in the same model complexity (Alvarez Melis and Jaakkola, 2018), so it becomes crucial to design models that reach a balance between these two elements, or moreover, close the gap between interpretable models and state-of-the-art models (Alvarez Melis and Jaakkola, 2018; Carvalho et al., 2019; Fan et al., 2021; Espinosa Zarlenga et al., 2022). We will discuss the detailed research efforts below:

- *Creating Transparent Reasoning Steps* In reasoning models, creating transparent minor steps is crucial to make the model interpretable (Hudson and Manning, 2018), and a number of papers accomplished it by introducing the MAC (Memory, Attention, and Composition) cell to separate memory and control (Hudson

and Manning, 2018), by utilizing other attention-based methods (Lin et al., 2019; Arik and Pfister, 2021), and by decomposing the complex reasoning process (Mascharka et al., 2018). These methods significantly improved the interpretability of the reasoning process but at the cost of model complexity and performance, though they close the gap of performance between interpretable and state-of-the-art models (Mascharka et al., 2018).

- *Distilling Complex Knowledge* Complex models, such as deep neural networks, often have high performance but lack transparency in their decision-making processes, making them difficult to interpret (Li et al., 2020). Knowledge distillation addresses this challenge by transferring knowledge from these complex, 'black-box' models (teachers) to simpler, more interpretable models (students). By introducing this structure into model design, student models can approximate the performance of the teachers while offering greater transparency, thus enhancing interpretability without sacrificing the capabilities of advanced machine learning models (Zhang et al., 2020b; Li et al., 2020). However, this interpretability is partial, especially in intricate missions, where the distilled knowledge may still be hard to interpret (Sachdeva and McAuley, 2023).

Moreover, the pronoun *Self-Explaining Models*, which can provide both prediction and explanation (Elton, 2020), was suggested by a number of papers as a better substitution to *Interpretable Models*, with many papers working on it (Alvarez Melis and Jaakkola, 2018; Rajagopal et al., 2021). For language models, the chain-of-thought (CoT) generation (Wei et al., 2022) may be recognized as a kind of self-explanation method.

4.2.2   Post Hoc Interpretability

This section explores techniques and methods applied to understand model internals after the models are trained and deployed, thus these techniques are often referred to as *post hoc interpretability* (Räuker et al., 2023). The goal is to understand the low-level structure and units of black-box neural networks and their causal effect on macroscopic behaviors and outputs.

Circuit Analysis   Circuits refer to the sub-networks within neural networks that can be assigned particular functionalities. As their counterparts in neuroscience, the neural circuits which are both anatomical and functional entities (Purves et al., 2001), circuits are also both physical and functional (Olah et al., 2020). Mechanistic interpretability researchers locate circuits in neural networks (microscopic) to understand model behaviors (macroscopic). Multiple circuits have been reported: curve circuits for curve detectors (OpenAI, 2021a), induction circuits for in-context learning (Olsson et al., 2022), indirect object identification circuits for identifying objects in sentences (Wang et al., 2022), Python docstrings for predicting repeated argument names in docstrings of Python functions (Heimersheim and Jett, 2023), grokking (Nanda et al., 2022), multi-digit addition (Nanda et al., 2022), and mathematical ability such as *greater than* (Hanna et al., 2024). Notably, many circuit analysis conducted to date has been focused on toy models and toy tasks (Räuker et al., 2023). The largest attempt to reverse engineer the natural behaviors of language models is finding the indirect object identification circuit, which is located in GPT-2 Small and has 28 heads (Wang et al., 2022).

Probing   Probing is a collection of techniques that train independent classifiers on the interested internal learned representations to extract concepts/features. One example is Gurnee et al used probing to study the linear representations of space and time in hidden layers. (Gurnee and Tegmark, 2023) Although probing has been favored by researchers to understand hidden layers (Alain and Bengio, 2017), it has limitations. For one, probing does help to understand learned representations in hidden layers, but it does not tell whether learned representations are used by models to produce predictions (Ravichander et al., 2021; Belinkov, 2022); for another, the issues of datasets may confound the issues with the model (Belinkov, 2022). In the context of safety and alignment, training probe requires the dataset to contain concepts/features of interest, which means probing can not be used to detect out-of-distribution features (i.e. features you suspect learned by the models but you don't have a dataset for them). Notably, representation engineering, built upon probing literature, is introduced to detect high-level cognitive phenomena and dangerous capabilities, including morality, emotion, lying, and power-seeking behaviors. (Zou et al., 2023a).

Dictionary learning   A key challenge of post hoc interpretability is *superposition*, *i.e.*, the tendency of neurons to encode more than one human-interpretable features simultaneously, which makes it very difficult to identify the individual features (Elhage et al., 2022). To address this challenge, methods based on dictionary learning has been proposed to separate these features in an unsupervised and scalable manner (Bricken et al., 2023).

Model Attribution   Attribution is a series of techniques that look at the contribution of some components (including head, neuron, layers, and inputs) for neuron responses and model outputs (Räuker et al., 2023). Gradient-based attribution is introduced to evaluate the quality of interpretation and guide the search for facts learned by the models (Ancona et al., 2018; Durrani et al., 2020; Lundstrom et al., 2022; Dai et al., 2022). However, those methods

are limited because they can not provide causal explanations (Räuker et al., 2023). Direct Logit Attribution is to identify the direct contribution of individual neurons to the prediction of the next neurons (Lieberum et al., 2023; McGrath et al., 2023; Belrose et al., 2023; Dar et al., 2023). But attribution methods also suffer from a salient constraint: they can only help with scenarios where you have datasets for features of interest. Consequently, such attribution methods cannot help with understanding out-of-distribution (OOD) features (including some misalignment scenarios) (Casper et al., 2023a).

Data attribution   Identifying the subset of training data that leads to a certain behavior can provide insight into both the safety of said behavior and ways to encourage or prevent that behavior. The method of *influence function* (Koh and Liang, 2017; Grosse et al., 2023) have been proposed to perform such attribution by approximating the result of leave-one-out training.

Visualization   Techniques of visualization help to understand neural structures, including techniques that visualize datasets (notably dimensionality reduction techniques) (Van der Maaten and Hinton, 2008; Olah, 2014, 2015), features (Erhan et al., 2009; Olah et al., 2017), weights, activations (Carter et al., 2019), structure (Reif et al., 2019), and the whole neural networks (Simonyan et al., 2013; Zeiler and Fergus, 2014; Nguyen et al., 2015; Karpathy et al., 2015; Mordvintsev et al., 2015; Nguyen et al., 2016; Kindermans et al., 2018). The purpose of visualization is to see neural networks with a new level of detail (Olah et al., 2020).

Perturbation and Ablation   These techniques are designed to test the counterfactual rather than the correlation (Räuker et al., 2023). Perturbation is a technique that modifies the input of models and observes changes in their outputs, and the ablation techniques knock out parts of neural networks[38], helping to establish a causal relationship between neural activation and the behavior of the whole network (Räuker et al., 2023).

Patching   Patching refers to the collection of methods *replacing* key components (paths and activations) and understanding counterfactual effects on model outputs. Among them, activations patching is a popular method among the safety community. Through applying activation patching and conducting both correct run and corrupted runs on the same neural network, researchers aim to locate key activations that matter more to the model output (Nanda, 2023a). In reality, patching is used to map and edit learning representations/concepts. Specific patching techniques include interpreting token representations in transformers (Li et al., 2021a; Bansal et al., 2021; Geva et al., 2021, 2022; Power et al., 2022; Olsson et al., 2022) and how do fully-connected layers learn these representations (Geva et al., 2021; Olsson et al., 2022), studying the key-query products to understand how do tokens attend to each other (Bahdanau et al., 2014; Lee et al., 2017; Liu et al., 2018; Strobelt et al., 2018; Clark et al., 2019; Vashishth et al., 2019; Vig, 2019; Hao et al., 2021; Chefer et al., 2021; Rigotti et al., 2022), identifying meaningful learned concepts from directions in latent space (from concepts to directions (Fong and Vedaldi, 2018; Kim et al., 2018), and from directions to post hoc explanations (Schneider and Vlachos, 2021)). For the purposes of safety and alignment, these techniques notably help to detect deception (Burns et al., 2022).

4.2.3   Outlook

Superposition makes the analysis at neuron level implausible   Superposition refers to the phenomenon that models represent more features than they have dimensions, so features would not correspond to neurons (Arora et al., 2018; Olah et al., 2020; Elhage et al., 2022). Superposition makes it hard to ensure AI safety by enumerating all features in a model (Elhage et al., 2022; Nanda, 2023b). Elhage et al. (2022) proposes three methods to solve superposition: creating models with no superposition (addressing it at training time), finding an overcomplete basis describing how features are stored in the neural nets (addressing it after the fact), or a mixture of both approaches. Notably, Bricken et al. (2023) builds a sparse auto-encoder to interpret group neurons, rather than individual neurons to extract features, which points out a promising direction to solve superposition: to move past it.[39]

Scalability   As is mentioned in the previous sections, there exists a trade-off between model interpretability and its capability (Alvarez Melis and Jaakkola, 2018), so interpreting real models while maintaining their performance will be harder than applying those techniques to toy models. Thus, scalability becomes a concern when interpretability researchers take a bottom-up approach to interpretability (mechanistic interpretability), as top-down methods such as attention mechanism (Hudson and Manning, 2018) would not face such a bottleneck. For mechanistic interpretability research, we either want to scale up techniques (*e.g.*, applying circuit analysis on real model (Wang et al., 2022)), or we want to scale up analysis (*e.g.*, finding larger structure in neural networks (Olah, 2023)). In the end, we want the microscopic analysis to answer the macroscopic model behavioral questions we care about (*e.g.*, in-context learning capability (Olsson et al., 2022) and more speculation about high-level cognitive capabilities such as planning and dangerous capability such as deception (Anthropic, 2023b)).

---

[38]Neurons (Zhou et al., 2018) and Subspace (Morcos et al., 2018; Ravfogel et al., 2022)

[39]see Elhage et al. (2022) for details on conceptual and empirical research questions about superposition
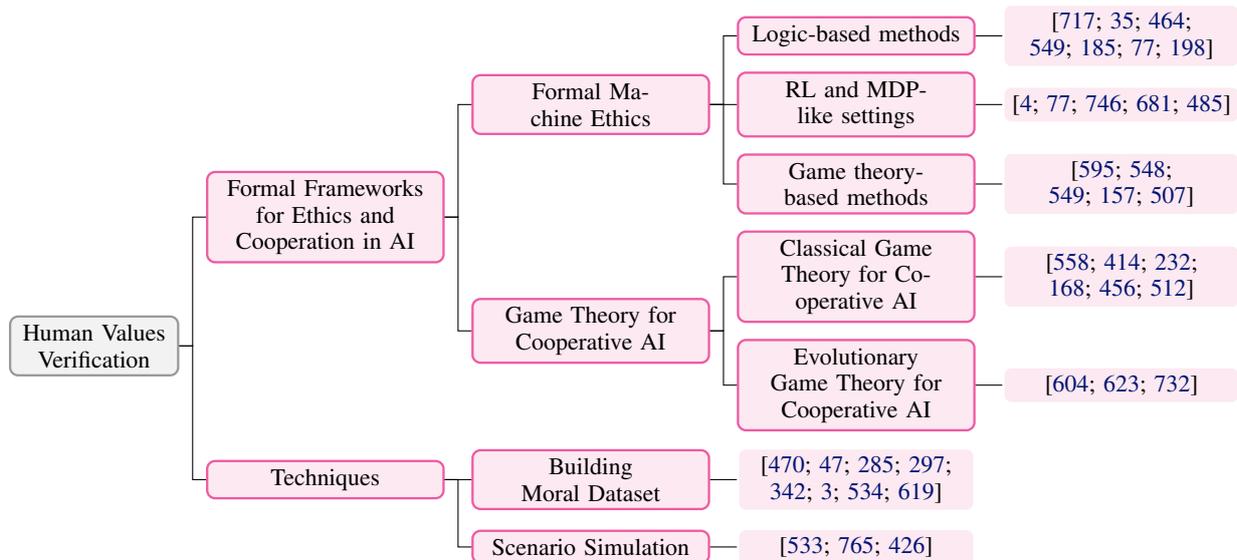
Figure 12: A Tree diagram summarizing the key concepts, logic, and literature related to Human Value Verification. The root of the tree represents Human Value Verification, which aims to *verify whether AI systems can adhere to the social norms and moral values*. The main branches represent the main structure of human value verification, including Formal Frameworks for Ethics and Cooperation in AI and specific Techniques of value verification. Further sub-branches list key works exploring each of these branches. This diagram provides an overview of research directions and specific techniques for making AI systems align with human values and social norms.

Evaluation and Benchmarking   Benchmarking offers insights about what methods work and quantifies their efficiency, and it will also drive community efforts in clear and meaningful directions (Lipton, 2018; Casper, 2023; Krishnan, 2020; Mohseni et al., 2021; Madsen et al., 2022). Interpretability benchmarks and metrics were made to evaluate interpretability tools (by evaluating their effectiveness in detecting trojans) (Casper et al., 2023a), circuits (by testing whether specific subgraphs are counted as circuits) (Lawrence et al., 2023) and explanations (by examining the faithfulness, comprehensiveness, and sufficiency of an explanation) (Lage et al., 2019; DeYoung et al., 2020; Krishna et al., 2022). However, as the inner logic of a certain AI system is unknown before the interpretability tools are applied (Samek et al., 2019) and different explanations may even contradict each other (Neely et al., 2021; Krishna et al., 2022), building a reliable evaluation benchmark or metric is rather difficult (Krishna et al., 2022).

4.3   Human Values Verification

Human Values Alignment refers to the expectation that AI systems should adhere to the community's social and moral norms (Chatila and Havens, 2019). As the capabilities of AI systems advance, some have begun to exhibit abilities approaching AGI (OpenAI, 2023a). In the future, we can expect autonomous agents governed by these AI systems to become an integral part of our daily lives (Lee et al., 2023b). However, if these systems fail to grasp the inherent complexity and adaptability of human values, their decisions could result in negative social outcomes. In this context, simply aligning with human intent may not be sufficient. Thus evaluating the alignment of human morality and values between AI systems and human beings becomes crucial (Weidinger et al., 2023). This underscores the importance of designing AI entities that are more socially oriented, reliable, and trustworthy. Following the logic of theoretical research and practical techniques, we divide our discussion of human value alignment into these two aspects: *Formulations* §4.3.1 and *Evaluation Methods* §4.3.2 of human value alignment.

4.3.1   Formulations

As the formulation of value is complicated, we introduce frameworks that formally characterize aspects of human values that are relevant to alignment. Specifically, we focus on two topics: *formal machine ethics* and *game theory for cooperative AI*. The former focuses on building a formal framework of machine ethics, while the latter discusses the value of multiagent systems, which share a similar origin of the game process.

Formal Machine Ethics   Machine ethics (Yu et al., 2018; Winfield et al., 2019; Tolmeijer et al., 2020), first introduced in §1.2.3, aim to build ethically-compliant AI systems. Here, we introduce the branch of machine ethics that focuses on formal frameworks – what we call *formal machine ethics*. We explain three approaches

to formal machine ethics: logic-based, RL/MDP-based, and methods based on game theory/computational social choice:

- **Logic-based methods**. One major direction within formal machine ethics focuses on logic (Pereira et al., 2016b). A number of logic-based works use or propose special-purpose logic systems tailored for machine ethics, such as the Agent-Deed-Consequence (ADC) model (Dubljevic, 2020), deontic logic (Von Wright, 1951; Arkoudas et al., 2005), event calculus and its variants (Berreby et al., 2017). Other works also develop methods for the formal verification of moral properties or frameworks for AI systems that accommodate such kind of formal verification (Dennis et al., 2016; Mermet and Simon, 2016).

- **RL & MDP-like settings**. Another line of work concerns statistical RL or other similar methods for planning within MDP-like environments (Abel et al., 2016; Svegliato et al., 2021). In particular, some works (Wu and Lin, 2018; Svegliato et al., 2021) involve the utilization of the manual design of ethics-oriented reward functions, a concept denoted as *ethics shaping*. Conversely, in other works (Berreby et al., 2017; Murtarelli et al., 2021), the segregation of ethical decision-making from the reward function is pursued.

- **Game theory-based methods**. To address multi-agent challenges, researchers have developed machine ethics methods based on game theory and computational social choice. Championed by Pereira et al. (2016a), methodologies of existing work can be broadly partitioned into Evolutionary Game Theory (EGT) (Pereira et al., 2016b), classical game theory (Conitzer et al., 2017), and computational social choice (Rossi et al., 2011; Noothigattu et al., 2018).

Game Theory for Cooperative AI   *Cooperative AI* (Dafoe et al., 2020, 2021) aims to address uncooperative and collectively harmful behaviors from AI systems (see §1.1.2). Here we introduce the branch of cooperative AI that focuses on game theory to complement the introduction to MARL-based cooperative training in §3.3.2. This branch tends to study the *incentives* of cooperation and try to enhance them, in contrast to the MARL's tendency to emphasize the *capabilities* of coordination. Examples of incentive failures include game theory dilemmas like the prisoner's dilemma (Phelps and Russell, 2023) and tragedy of the commons (Perolat et al., 2017), while examples of coordination capability failures include bad coordination of a robot football team (Ma et al., 2022).

- **Classical Game Theory for Cooperative AI.** A number of works focus on classical game theory as a setting for cooperative AI. Among them, one salient theme is that of *Stackelberg games*, *i.e.* games where one player (the "leader") moves first, and all other players (the "followers") move in response to the leader's move. This is suitable for modeling *commitment* in games (*i.e.*, a player pre-committing to a certain move or strategy to gain an advantage), and, according to Dafoe et al. (2020), understanding commitment is one of the four pillars of cooperative AI research. Recent works on Stackelberg games include the introduction of bounded rationality into the model (Pita et al., 2010), dynamic models (Li and Sethi, 2017), machine learning of Stackelberg equilibria (Fiez et al., 2020), and more. Apart from Stackelberg games, Dafoe et al. (2020) has highlighted the importance of studying *mixed-motive games* (*i.e.*, general games that are neither purely cooperative nor purely competitive) due to their realisticity. Examples of recent work on this front include McKee et al. (2020), which finds a positive correlation between values diversity in synthetic populations and performance in mixed-motive games, and Oesterheld and Conitzer (2022), which constructs interventions on the payoff matrix of general games to induce Pareto improvements in game outcome.

- **Evolutionary Game Theory for Cooperative AI**. Another avenue of research, initiated by Sachs et al. (2004), aims to understand how cooperation emerges from evolution – this includes human cooperation, which arose from Darwinian evolution, as well as the cooperation tendencies in AI systems that could emerge within other evolutionary settings such as the replicator dynamics (Schuster and Sigmund, 1983). These works adopt a methodology called *evolutionary game theory* (Weibull, 1997), which studies, often using tools from dynamical systems, the long-run evolutionary outcome of a large population of agents whose reproductive success is determined by game outcomes against others. More recent work on this front tends to add features to the model to improve its realisticity, including, for example, population structures and complexity costs on strategies.

4.3.2   Evaluation Methods

In this section, we assume that we have already obtained the appropriate value that should be aligned. However, even so, under the guidance of Goodhart's Law (Goodhart and Goodhart, 1984), we cannot simply define complex human values as reward functions, which also brings greater challenges to value alignment. We introduce specific human value alignment techniques in three parts: *Building Moral Dataset*, *Scenario Simulation*.

Building Moral Dataset *Moral Alignment* refers to the adherence of AI systems to human-compatible moral standards and ethical guidelines while executing tasks or assisting in human decision-making (Min et al., 2023). Early attempts at moral value alignment, initiated in 2018 (Awad et al., 2018), have confirmed that the definition and evaluation of moral values themselves is a challenging issue. This has led to the emergence of abstract moral standards (Hagendorff, 2022) and various different standards driven by the average values of diverse community groups (Awad et al., 2018), fueling further in-depth research into moral value assurance.

Assurance of moral values is typically achieved by constructing corresponding datasets. The Rule-of-Thumb (RoT) serves as a gauge for determining what actions are considered acceptable in human society. Building on this concept, Emelin et al. (2021); Forbes et al. (2020); Ziems et al. (2022) introduced the Moral Stories, SOCIAL-CHEM-101, and Moral Integrity Corpus datasets respectively, focusing on providing human social and moral norms. Hendrycks et al. (2020) and Jin et al. (2022) introduced the ETHICS and MoralExceptQA datasets respectively, highlighting the inability of contemporary models to align ethically with human values. Abdulhai et al. (2022) found that models exhibit certain morals and values more frequently than others, revealing how the moral foundations demonstrated by these models relate to human moral foundations. Pan et al. (2023b) explored the trade-off between rewards and moral behavior, discovering a certain tension between the two.

Scenario Simulation *Scenario simulation* is a more complex form than datasets and therefore is considered by some views to be more effective in replicating real situations and harvesting better results. The form of the scenario can also vary. Pan et al. (2023a) built a series of diverse, morally salient scenarios through text adventure games, evaluating complex behaviors such as deception, manipulation, and betrayal. On the other hand, some work attempts to make intelligent agents learn human values through simulating human-machine interaction. Yuan et al. (2022) proposed a method for bidirectional value alignment between humans and machines, enabling machines to learn human preferences and implicit objectives through human feedback. Liu et al. (2024a) placed AI within a simulated human society sandbox, allowing AI to learn human societal value inclinations by mimicking human-social interactions.

5 Governance

Besides technical solutions, governance, the creation and enforcement of rules, is necessary to ensure the safe development and deployment of AI systems. In this section, we survey the literature on AI governance by exploring the role of AI governance, the functions, and relationships between stakeholders in governing AI, and several open challenges to effective AI governance.

5.1 The Role of AI Governance

To explore the role of AI governance, we must identify the challenges that require governance. A range of social and ethical issues can and have already emerged from the adoption and integration of AI into various sectors of our society (AI Safety Summit, 2023). For instance, AI applications can inadvertently perpetuate societal biases, resulting in racial and gender discrimination (Caliskan et al., 2017; Perez et al., 2023). Moreover, unchecked reliance on these systems can lead to repercussions such as labor displacement (Acemoglu and Restrepo, 2018), widening socioeconomic disparities, and the creation of monopolistic environments.

AI systems have exhibited the potential to jeopardize global security (Turchin and Denkenberger, 2020). For example, OpenAI's system card for GPT-4 (OpenAI, 2023a) finds that an early version of the GPT-4 model as well as a version fine-tuned for increased helpfulness and harmlessness exhibits capabilities to enable disinformation, influence operations, and engineer new biochemical substances, among other risky behavior. Urbina et al. (2022) further demonstrated the potential of AI systems to enable the misuse of synthetic biology by inverting their drug discovery model to produce 40,000 toxic molecules.

The horizon also holds the prospect of increasingly agentic and general-purpose AI systems that, without sufficient safeguards, could pose catastrophic or even existential risks to humanity (McLean et al., 2023). For example, OpenAI's Weng (2023b) argued that models such as LLM could essentially act as the brain of an intelligent agent, enhanced by planning, reflection, memory, and tool use. Projects such as AutoGPT, GPT-Engineer, and BabyAGI epitomize this evolution. These systems can autonomously break down intricate tasks into subtasks and make decisions without human intervention. Microsoft research suggests that GPT-4, for instance, hints at the early inklings of AGI (Bubeck et al., 2023). As these systems evolve, they might lead to broad socio-economic impacts such as unemployment, and potentially equip malicious actors with tools for harmful activities.

The major objective of AI governance is to mitigate this diverse array of risks. In pursuit of this goal, relevant actors should maintain a balanced portfolio of efforts, giving each risk category its due consideration.

5.2 The Multi-Stakeholder Approach

We put forward a framework to analyze the functions and relationships between stakeholders in AI governance (see Figure 13). In this framework, we outline three main entities. **Government Agencies** oversee AI policies
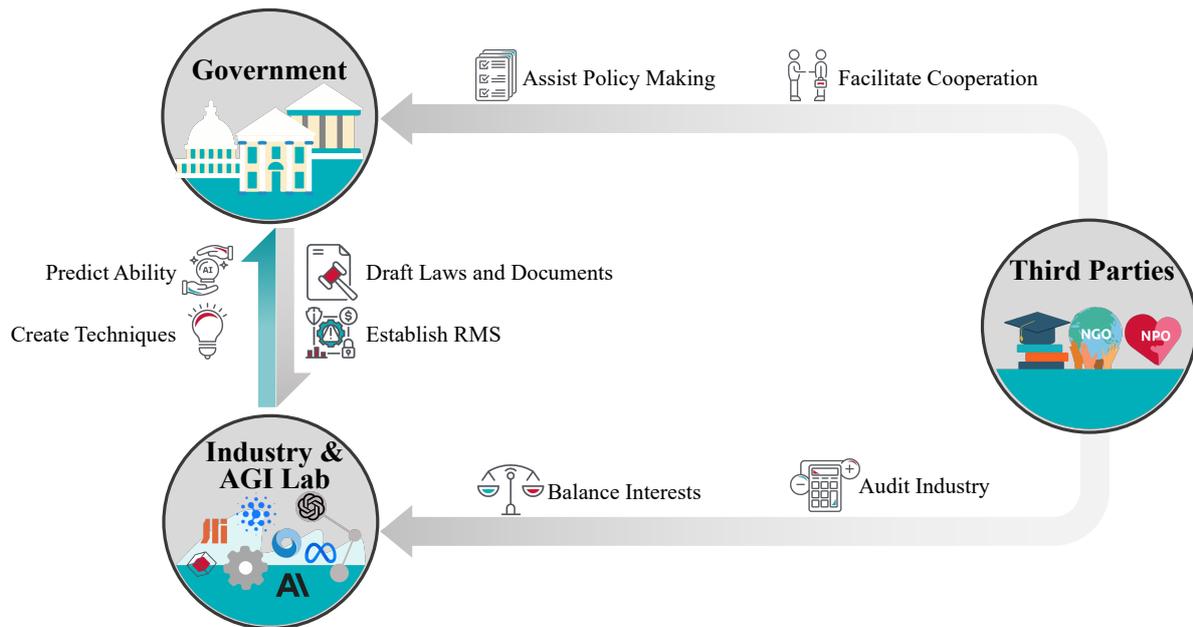
Figure 13: Our framework for analyzing AI governance at present. The proposed framework explains the nonexhaustive interrelationships and functions among three primary entities in AI governance: the government, industry and AGI labs, and third parties. The government's governance role encompasses regulating the industry and AGI labs and directing the trajectory of future AI development through policy documents. It also devises a *Risk Management System* (RMS) (Mannes, 2020) to abate AI-related threats. Industry and AGI labs return by offering watchful predictions into AI development and innovating new technological methodologies to support regulatory measures (such as model evaluation (Shevlane et al., 2023)). Third parties fulfill a dual function, offering expert advice for robust governmental policy development and fostering collaborations among governments. In the context of industry and AGI labs, these third parties assist in equilibrating corporate interests to prevent disorganized competition from information asymmetry. They also deliver auditing services to the industry and AGI labs as independent entities.

using legislative, judicial, and enforcement powers, as well as engage in international cooperation. **Industry and AGI Labs** research and deploy AI technologies, making them subjects of the governance framework, while proposing techniques to govern themselves and affecting governance policy. **Third Parties**, including academia, Non-Governmental Organizations (NGOs), and Non-Profit Organizations (NPOs), perform not only auditing on corporate governance, AI systems, and their applications but also assist the government in policy-making.

Proposals have been made about specific principles for a multi-stakeholder AI governance landscape. Notably, Brundage et al. (2020) argues to implement institutions, software, and hardware to make claims about the safety of AI systems more verifiable.

Government   According to Anderljung et al. (2023), three building blocks for government regulation are needed: (1) standard development processes to determine appropriate requirements for cutting-edge AI developers, (2) registration and reporting requirements to offer regulators insight into the progress of advanced AI development processes, (3) mechanisms to guarantee adherence to safety standards in the development and deployment of cutting-edge AI models.

At present, an emerging collection of governmental regulations and laws is surfacing on a global scale, including the *European Union's AI Act* (European Parliament, 2023), and the *Bipartisan Framework for U.S. AI Act* (Blumenthal and Hawley, 2023). Such regulations are indispensable for the safety and alignment of AI systems.

Industry and AGI Labs   Governance efforts in industry and AGI labs should emphasize comprehensive AI risk assessments throughout the lifecycle of the AI system. Based on discussions in Koessler and Schuett (2023); Schuett et al. (2023), the full cycle of AI risk assessment can be seen as consisting of five stages. **Pre-development risk assessments**, **pre-training risk assessments**, and **pre-deployment risk assessments** all include predictions and analyses of impact and risks with a variety of tools, but with increasing amounts of detail, clarity, and sophistication (Koessler and Schuett, 2023). **Post-deployment monitoring** is the phase where mechanisms for monitoring are established, and all previous analyses are continually updated post-deployment (Koessler and Schuett, 2023). **External scrutiny** includes bug bounty programs (Schuett et al., 2023), external red teaming and third-party model

auditing (Schuett et al., 2023; Anderljung et al., 2023)

Taking security measures against the risks associated with AI systems seems to be widely accepted by AI companies and related practitioners. Schuett et al. (2023) shows that 98% of respondents who have been surveyed somewhat or strongly approved that AGI labs should perform pre-deployment risk assessments, hazardous capabilities evaluations, third-party model audits, safety restrictions on model usage, and red teaming to guarantee AI safety. Meanwhile, leading AI companies, including Amazon, Anthropic, Google, Inflection, Meta, Microsoft, and OpenAI, have voluntarily committed to the government to implement security measures (The White House, 2023).

Notably, a lot of researchers have proposed pausing the development of advanced AI systems to win more time for safety research, risk assessments, and regulatory preparations (Bengio et al., 2023). Their proposals include blanket pausing of all sufficiently advanced systems (Bengio et al., 2023), and also conditional pausing of specific classes of models in response to evaluation results on specific failure modes (Alaga and Schuett, 2023), including the currently adopted practice of *responsible scaling policy* (RSP) (Anthropic, 2023a).

Third Parties   Mökander et al. (2023) presents three key functions of third-party auditing: (1) *Governance audits* (of tech providers that design and disseminate LLMs) (2) *Model audits* (of LLMs after pre-training but prior to their release) (3) *Application audits* (of applications based on LLMs).

One prominent example of existing third-party audits is that of METR, initially a project of Alignment Research Center (ARC Evals, 2023; Kinniment et al., 2023), who collaborated with OpenAI to perform red teaming on GPT-4 (OpenAI, 2023a) and partnered with Anthropic to perform red teaming on Claude 2 (Anthropic, 2023c). These efforts include evaluations on toxicity and bias, as well as frontier AI risks such as autonomous replication, manipulation, cybersecurity, and biological weapon risks (OpenAI, 2023a; Shevlane et al., 2023).

Apart from auditing, third parties can support AI governance in other ways, such as assisting policy-making and facilitating cooperation internationally (Ho et al., 2023). For example, Maas (2021) thinks that the government should prefer technology-neutral rules rather than technology-specific rules. *AI4People's Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations* (Floridi et al., 2021), released by AI4People, was guided to the Ethics Guidelines for Trustworthy Artificial Intelligence presented in April 2019 (Atomium-EISMD, 2023). The World Economic Forum (WEF) convenes government officials, cooperations, and civil society and it has initiated a Global AI Action Alliance in collaboration with partner organizations, with the goal of promoting international cooperation in the field of AI (Kerry et al., 2021).

5.3   Open Problems

There are numerous open problems in the existing field of AI governance. These problems often have no clear answers, and discussion of these questions can often promote better governance. For effective AI governance, we mainly discuss international governance and open-source governance, hoping to promote the safe development of AI through our discussion.

5.3.1   International Governance

Amidst the swift progress and widespread implementation of AI technology universally, the need for international governance of AI is high on the agenda (Summit, 2023). Critical discussions revolve around the necessity to institute a global framework for AI governance, the means to ensure its normativity and legitimacy (Erman and Furendal, 2022), among other significant concerns. These themes draw an intensifying level of detail and complexity in their consideration. Also, as stated by United Nations secretary-general António Guterres during a Security Council assembly in July, generative AI possesses vast potential for both positive and negative impacts at scale, and failing to take action to mitigate the AI risks would be a grave neglect of our duty to safeguard the well-being of current and future generations (Guterres, 2023), international governance also has intergenerational influence. Hence, we examine the significance and viability of international AI governance from three aspects within this section: *manage global catastrophic AI risks*, *manage opportunities in AI*, and *historical and present efforts*, with both generational and intergenerational perspectives. We aim to contribute innovative thoughts for the prospective structure of international AI governance.

Manage Global Catastrophic AI Risks   The continual advancements in AI technology promise immense potential for global development and prosperity (Vinuesa et al., 2020). However, they inevitably harbor underlying risks. The unchecked competition in the market and geopolitical factors could precipitate the untimely development and deployment of advanced AI systems, resulting in negative global externalities (Tallberg et al., 2023). The amplification of existing inequalities such as racial and gender bias (Swaugerarchive, 2020) ingrained in AI systems may result in intergenerational ethical discrimination. Since these risks are international and intergenerational, it seems that international governance interventions could alleviate these catastrophic global AI challenges. For example, a consensus amongst nations could help defuse potential AI arms races, while an industry-wide agreement could avert the hasty and irresponsible development of sophisticated AI systems, thus securing the long-term and sustainable development of AI (Ho et al., 2023).

Manage Opportunities in AI   The opportunities created by AI development are not distributed equally, which may cause enduring digital inequality between regions and harm the sustainability of AI development. Geographic variances in AI progression suggest an inequitable distribution of its economic and societal benefits, potentially excluding developing nations or specific groups from these advantages (Ho et al., 2023; Tallberg et al., 2023). Moreover, the consolidation of decision-making authority within the technology sector among a limited number of individuals (Sara Stratton, 2021; Noble et al., 2021) could cause an intergenerational impact. Such inequality in the distribution of interests can be mitigated through international governance. Effective international consensus and coordination on the allocation of AI opportunities, which is facilitated by its propagation, education, and infrastructural development (Opp, 2023), could ensure a balanced distribution of benefits derived from AI and promote sustainability in its ongoing development.

Historical and Present Efforts   Before the surge of AI technology, the international community had laid down frameworks in line with cooperative regulation of influential technologies and critical matters. For example, the Intergovernmental Panel on Climate Change (IPCC) convened specialists to assess climactic environmental issues, fostering scientific consensus (Ho et al., 2023). The International Civil Aviation Organization (ICAO) standardized and oversaw international regulations, simultaneously assessing the member nations' compliance with these laws (Ho et al., 2023). The International Atomic Energy Agency (IAEA) propelled the harmonious utilization of nuclear energy, with its global reach and sophisticated monitoring and evaluation mechanisms. Fast forward to the present-day scenario, wherein multiple international organizations have arrived at a consensus on AI governance. In 2019, the G20 members consolidated a ministerial declaration focusing on human-centered artificial intelligence principles (G20, 2019). Concurrently, the Organisation for Economic Cooperation and Development (OECD) set forth the *OECD Principles on Artificial Intelligence* (OECD, 2019). The IEEE Standards Association launched a worldwide initiative aimed at *Securing that all stakeholders involved in the design and implementation of autonomous and intelligent systems receive proper education, training, and motivation to emphasize ethical concerns, thereby advancing these technologies for the betterment of humanity.* (Chatila and Havens, 2019). In 2021, the United Nations Educational, Scientific and Cultural Organization(UNESCO) produced the first-ever global standard on AI ethics (UNESCO, 2021), which aims to lay the foundations for making AI systems work for the good of humanity and societies, and to prevent potential harm caused by losing control over AI systems. In 2023, the AI Safety Summit was convened in London, United Kingdom. Countries held roundtable discussions on the risks and opportunities of AI and jointly issued the Bletchley Declaration (Summit, 2023). The scholarly community has also proposed prospective international governance frameworks for AI, such as the International AI Organization (IAIO) (Trager et al., 2023). We hope these precedents and research outcomes will inspire and provide the groundwork for developing a resilient and long-lasting international framework for AI governance in the future.

5.3.2   Open-Source Governance

The debate over the open-sourcing of contemporary AI models is contentious in AI governance, particularly as these models gain increased potency (Seger et al., 2023). The potential security hazards linked with making these models open-source continue to be the crux of debates among AI researchers and policymakers. The offence-defence balance in open-source AI governance also remains controversial. There is still debate over whether open-source models will increase model security or increase the risk of abuse. As referenced in Shapiro and Siegel (2010), the efficacy of disclosure depends on the chance of potential attackers already possessing the knowledge, coupled with the government's capacity to convert transparency into the identification and solution of emerging vulnerabilities. Some scholars have already conducted preliminary discussions on the offense-defense balance in the AI field, such as Weng (2023a)'s discussion of adversarial attacks. If a suitable equilibrium between offence and defence cannot be forged for AI systems, the open-sourcing could potentially give rise to significant risks of AI system misuse.

For precision and clarity, we adhere to the definition of open-source models stated by Seger et al. (2023): enabling open and public access to the model's architecture and weights, allowing for modification, study, further development, and utilization by anyone. Currently, the most recognized open-source models include Llama2, Falcon, Vicuna, and others. In this section, we evaluate the security advantages and potential threats posed by open-source models, fostering a discourse on the feasibility of open-sourcing these models. Ultimately, our objective is to amalgamate insights from existing studies to put forward suggestions for future open-source methodologies that will ascertain the secure implementation of these models.

Arguments for Open-sourcing   The view that supports the open-sourcing of existing models suggests that this method can mitigate the security risks inherent in these models in several ways:

- **Potentially Bolster Model's Security**. Meta's assertions in their release blog for Llama2 (Meta, 2023) promote the belief that this enables the developer and the technical community to conduct tests on the models.

As a result, this rapid identification and resolution of issues can considerably strengthen model security. In contrast, another perspective suggests that open-sourcing existing models could enhance the recognition of associated risks, thereby facilitating a greater focus on, investigation into, and mitigation of these potential hazards (Zellers, 2019).

- **Foster the Decentralization of Power and Control**. Open-sourcing has been widely recognized as an effective strategy in reducing the dominance of major AI laboratories across various sectors, including economic, social, and political domains (Seger et al., 2023). An example is articulated in the core reasons for Stability's open-sourcing of Stable Diffusion: They place their trust in individuals and the community, as opposed to having a centralized, unelected entity controlling AI technology (Mostaque, 2022). Furthermore, certain commentators draw an analogy between open-sourcing and the Enlightenment Era, asserting that decentralized control reinforces faith in the power and good of humanity and society (Howard, 2023), implementing central regulations for safety purposes might amplify the power of the AI technology community instead.

Arguments against Open-sourcing   Critics of open-source models assess the potential for misuse from the following viewpoints:

- **Potentially Be Fine-Tuned into Detrimental Instances**. Current research rigorously affirms that AI systems, contradictory to their initial design intent for mitigating toxicities in chemistry or biology, now hold the potential to manufacture new chemical toxins (Urbina et al., 2022) and biological weaponry (Sandbrink, 2023). The malicious fine-tuning of such models could lead to profound security risk manifestations. Besides, language models, once fine-tuned, could emulate skilled writers and produce convincing disinformation, which may generate considerable sociopolitical risks (Goldstein et al., 2023).

- **Inadvertently Encourage System Jailbreaks**. Research indicates that unfettered access to open-sourced model weights could facilitate bypassing system security measures (Seger et al., 2023). This premise was epitomized by Zou et al. (2023b), who showcased this potentiality by developing attack suffixes using Vicuna-7B and 13B. Once implemented within readily accessible interfaces such as ChatGPT, Bard, and Claude, these provoked unwanted generations. Therefore, open-sourcing a model might unintentionally undermine the safeguarding protocols of models that are not open-sourced, consequently amplifying the likelihood of model misuse.

Tentative Conclusions on Open-Source Governance   The debate on the open-sourcing of AI models remains unsettled, with a prevailing viewpoint that the disclosure of AI models does not pose significant risks at present. Our discourse not only synthesizes existing perspectives on this topic but also prepares the ground for future deliberations considering the prudence of open-sourcing more advanced AI systems.

Existing guidelines for open-sourcing advanced AI systems include measures such as evaluating risks by quantifying the potential for misuse via fine-tuning and a gradual model release (Solaiman et al., 2019; Seger et al., 2023). Meanwhile, policymakers are establishing rigorous compliance protocols for these open-source models. For example, European policymakers insist that the models should have "performance, predictability, interpretability, corrigibility, security, and cybersecurity throughout [their] lifecycle." (Chavez, 2023).

5.4   Rethinking AI Alignment from Socio-technical Perspective

In the preceding discussion, our primary focus is on AI systems as the core of AI Alignment. We examine strategies to align the system with human intentions and values throughout its lifecycle, considering both forward and backward alignment. In the future, AI will address more challenging and high-stakes decisions, *e.g.*, "How to allocate resource for fairness?" and "Which drugs are safe to approve?". These decisions will require not only significant expertise for well-informed answers but also involve value judgments, leading to strong disagreements among informed individuals based on differing values. Furthermore, AI systems may transmit incorrect values, sway public opinion, facilitate cultural invasion, and exacerbate social division (Goldstein et al., 2023). Singapore Conference on AI (SCAI) once introduced 12 questions that are meant to be a holistic formulation of the challenges that should be addressed by the global AI community to allow humanity to flourish [40]. In the area of alignment we are more concerned about the following question: as AI systems evolve into socio-technical entities, how can alignment techniques mit igate the challenges they pose to human society? Specifically, we explore the incorporation of values into AI systems through alignment techniques and provide insights into security methods. We also aim to identify the alignment techniques needed to address the socio-technical challenges posed by future AI systems.

---

[40] https://www.scai.gov.sg/

5.4.1   How to incorporate value into AI systems?

Aligning AI systems with human morals and societal values is a key objective of alignment technology. However, current technologies (*e.g.*, RLHF) primarily blend preferences without distinguishing specific values, focusing solely on human preferences. Human preferences effectively address the basic alignment issue: ensuring models align with human intentions and safety, but not morals and societal values. However, minor errors in future AI systems' critical problems can lead to disagreements among people with differing viewpoints. Truly understanding human values is crucial for AI systems to generalize and adapt across various scenarios and ideologies. Incorporating values into AI systems generally involves two aspects: aligning with individual values (§4.3), and aligning with collective values.

In this part, we mainly discuss the second topic. The main challenge of collective value alignment lies in determining which groups to include. A prevalent approach is defining universal values like fairness, justice, and altruism, exemplified by the veil of ignorance. However, this work remains theoretical; another approach avoids defining universal values, instead seeking the broadest overlap of values across cultures. Bakker et al. (2022) initiated this approach by gathering preferences from various demographics, training a language model, and aggregating results using diverse social welfare functions. Similarly, simulated deliberative democracy has been proposed to enhance decision-making (Leike, 2022). Specifically, individuals from diverse demographics reach consensus on value-laden topics with AI assistance. This data informs new model training, enabling simulation of deliberative democracy for more apt responses to new value-laden issues.

Furthermore, instead of providing a consensus answer to all users, collective value alignment should encourage AI systems to tailor responses to specific demographic groups. In other words, what values should guide the model's responses to specific questions or in certain dialogues? Democratic Fine-Tuning (MAI, 2023) uses a value card and moral graph to link various values, allowing fine-tuned LLMs to reflect on their moral context before responding.

However, while most value discussions assume static values, social values are actually dynamic and evolving. Exploring how value-aligned AI systems can dynamically adapt to changing environmental values is crucial. Future technologies need to address static value alignment first, including strategies for sampling human groups for alignment. Bakker et al. (2022) founds that consensus statements built silently from a subgroup will lead to dissent among excluded members, highlighting the consensus's sensitivity to individual input. For international cooperation, establishing a shared data center is necessary but also requires first determining which civilizations to include and if their values can align.

5.4.2   How to use Alignment techniques to support AI Governance?

It's crucial to ensure the reliability and trustworthiness of AI systems as they are adopted in various real-world decision-making scenarios. On one hand, language models still exhibit illusions during use, and on the other hand, the reliability of systems comprises two parts: the system's reliability under individual testing environments and its reliability in human interactions. Another issue is constructing systems with decision-making processes that are observable and explainable to users. From a social perspective, the proliferation of AI systems across fields also poses potential risks. This risk arises from a gap between AI developers, who often focus on advancing technology without considering its downstream applications, and AI adopters, who may transfer AI systems to their fields without adequate safety considerations or verification of replicable success [41]. Therefore, it is crucial to build a framework that enables AI adopters to accurately assess model utility and appropriateness, and allows AI regulators to quickly identify risks and issue safety alerts in AI systems.

Alignment techniques can facilitate synchronized, independent, and rigorous evaluations of AI systems. AI developers should prioritize appropriate bias handling during the training process, acknowledging the importance of socio-economic, cultural, and other differences. Furthermore, we should aim to develop robust and fair evaluation methods and datasets for auditing AI systems. Zhu et al. (2023) proposes the first dynamic testing protocol for large language models, utilizing Directed Acyclic Graphs (DAGs) to dynamically generate test data, thereby reducing the risk of test data memorization and contamination. Additionally, new robust security protocol evaluation methods have been introduced: Shlegeris and Greenblatt (2023) suggests constructing adversarial policies to manage dangerously powerful and deceptive models, while Greenblatt et al. (2023) proposes (un)trusted editing to supervise models based on their harm and deceitfulness levels. Future efforts should also prevent AI systems from reward-hacking evaluation system exploits and aim to provide AI regulators with an explainable, independent, and centralized evaluation system.

AI adopters and the industry should allocate financial and computational resources to thoroughly evaluate use cases and share case studies showcasing both successes and failures. Equally important is training for adopters on downstream applications.

---

[41]https://www.scai.gov.sg/scai-question-11/

6 Conclusion

In this survey, we have provided a broadly-scoped introduction to AI alignment, which aims to build AI systems that behave in line with human intentions and values. We specify the objectives of alignment as Robustness, Interpretability, Controllability, and Ethicality (**RICE**), and characterize the scope of alignment methods as comprising of *forward alignment* (making AI systems aligned via alignment training) and *backward alignment* (gaining evidence of the systems' alignment and govern them appropriately to avoid exacerbating misalignment risks). Currently, the two notable areas of research within forward alignment are *learning from feedback* and *learning under distribution shift*, while backward alignment is comprised of *assurance* and *governance*.

One thing that sets alignment apart from many other fields is its diversity (Hendrycks, 2022) – it is a tight assembly of multiple research directions and methods, tied together by a shared goal, as opposed to a shared methodology. This diversity brings benefits. It fosters innovation by having the different directions compete and clash against each other, leading to a cross-pollination of ideas. It also allows different research directions to complement each other and together serve the goal of alignment; this is reflected in the *alignment cycle* (see Figure 2), where the four pillars are integrated into a self-improving loop that continually improves the alignment of AI systems. Meanwhile, this diversity of research directions raises the barrier to entry into this field, which mandates the compilation of well-organized survey materials that serve both the newcomers and the experienced. In this survey, we attempt to address this need by providing a comprehensive and up-to-date overview of alignment.

We attempt to account for the full diversity within the field by adopting a broad and inclusive characterization of alignment. Our survey of alignment gives a spotlight to almost all major research agendas in this field, as well as to real-world practices on the assurance and governance front. We recognize that boundaries of alignment are often vague and subject to debate. Therefore, when proposing the RICE principles, we put forth our broad characterization of alignment as an explicit choice. In the meantime, we recognize that such a survey needs to be a long-term endeavor that is continually reviewed and updated. Both the problems and methods of alignment closely follow the development of machine learning. This fast-paced development means that new materials and frameworks can become outdated after merely a few years. This fact is one reason why we write the survey to reflect the latest developments, and also mandates continual maintenance and updates.

We conclude the survey by looking ahead and presenting the key traits in this field that we believe ought to be preserved or fostered.

Open-Ended Exploration of Novel Challenges and Approaches   A lot of the alignment discourse is built upon classic works that predate the recent developments of LLMs and other breakthroughs in large-scale deep learning. Thus, when this paradigm shift happens in the machine learning field, it is plausible that some challenges in alignment become less salient while others become more so; after all, one defining feature of scientific theories is their falsifiability (Popper, 2005). More importantly, this shift in machine learning methodology and the broader trend of ever-tighter integration of AI systems into society (Abbass, 2019) introduces novel challenges that could not be envisioned before. This requires that we engage in *open-ended exploration*, actively seeking out new challenges that were previously neglected. Moreover, such an exploration need not be constrained to challenges – a similar mindset should be adopted regarding approaches and solutions, thus building a more diverse portfolio for both the *questions* and the *answers* (Shimi, 2022).

Combining Forward-Looking and Present-Oriented Perspectives   Alignment has emphasized harms from potential advanced AI systems that possess stronger capabilities than current systems (Ngo, 2020a). These systems might come into existence well into the future, or might just be a few years away (Stein-Perlman et al., 2022). The former possibility requires us to look into extrapolated trends and hypothetical scenarios (Carlsmith, 2022). In contrast, the latter possibility highlights the need for on-the-ground efforts that work with current governance institutions and use current systems as a prototype for more advanced ones (Cotra, 2021).

Emphasis on Policy Relevance   Alignment research does not live in a vacuum but in an ecosystem ,[42] with participation from researchers, industry actors, governments, and non-governmental organizations. Research serving the needs of the AI alignment and safety ecosystem would therefore be useful. Such needs include solving the key barriers to various governance schemes, for example, extreme risk evaluations (Shevlane et al., 2023), infrastructure for computing governance, and mechanisms for making verifiable claims about AI systems (Brundage et al., 2020).

Emphasis on Social Complexities and Moral Values   As AI systems become increasingly integrated into society (Abbass, 2019), alignment ceases to be only a single-agent problem and becomes a social problem. Here, the meaning of *social* is three-fold.

---

[42]See `aisafety.world` for a map of the organizational landscape of alignment.

1. Alignment research in multi-agent settings featuring the interactions between multiple AI systems and multiple humans (Critch and Krueger, 2020).

2. Incorporating human moral and social values into alignment (see §1.2.3 and §4.3), which is closely linked to the field of *machine ethics* and *value alignment* (Gabriel, 2020; Gabriel and Ghazavi, 2021).

3. Modeling and predicting the impacts of AI systems on society, which requires methods to approach the complexities of the social system, including those from the social sciences. Examples of potentially useful methodologies include social simulation (Bonabeau, 2002; De Marchi and Page, 2014; Park et al., 2023a) and game theory (Critch and Krueger, 2020).

Acknowledgments

References

[1] Hussein A Abbass. 2019. Social integration of artificial intelligence: functions, automation allocation logic and human-autonomy trust. *Cognitive Computation*, 11(2):159–171.

[2] Pieter Abbeel and Andrew Y Ng. 2004. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 1.

[3] Marwa Abdulhai, Clément Crepy, Daria Valter, John Canny, and Natasha Jaques. 2022. Moral foundations of large language models. In *AAAI 2023 Workshop on Representation Learning for Responsible Human-Centric AI*.

[4] David Abel, James MacGlashan, and Michael L Littman. 2016. Reinforcement learning as a framework for ethical decision making. In *AAAI Workshop: AI, Ethics, and Society*, volume 16, page 02. Phoenix, AZ.

[5] Daron Acemoglu and Pascual Restrepo. 2018. Artificial intelligence, automation, and work. In *The economics of artificial intelligence: An agenda*, pages 197–236. University of Chicago Press.

[6] Stephen Adams, Tyler Cody, and Peter A Beling. 2022. A survey of inverse reinforcement learning. *Artificial Intelligence Review*, 55(6):4307–4346.

[7] Gediminas Adomavicius, Jesse Bockstedt, Shawn Curley, and Jingjing Zhang. 2022. Recommender systems, ground truth, and preference pollution. *AI Magazine*, 43(2):177–189.

[8] M Mehdi Afsar, Trafford Crump, and Behrouz Far. 2022. Reinforcement learning based recommender systems: A survey. *ACM Computing Surveys (CSUR)*, 55(7):1–38.

[9] Forest Agostinelli, Guillaume Hocquet, Sameer Singh, and Pierre Baldi. 2018. From reinforcement learning to deep reinforcement learning: An overview. In *Braverman Readings in Machine Learning. Key Ideas from Inception to Current State: International Conference Commemorating the 40th Anniversary of Emmanuil Braverman's Decease, Boston, MA, USA, April 28-30, 2017, Invited Talks*, pages 298–328. Springer.

[10] AI Safety Summit. 2023. Ai safety summit 2023: Roundtable chairs' summaries, 1 november. https://www.gov.uk/government/publications/ai-safety-summit-1-november-roundtable-chairs-summaries/ai-safety-summit-2023-roundtable-chairs-summaries-1-november--2.

[11] Ajeya Cotra. 2021. why-ai-alignment-could-be-hard-with-modern-deep-learning. https://www.cold-takes.com/why-ai-alignment-could-be-hard-with-modern-deep-learning.

[12] Riad Akrour, Marc Schoenauer, and Michele Sebag. 2011. Preference-based policy learning. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011. Proceedings, Part I 11*, pages 12–27. Springer.

[13] Riad Akrour, Marc Schoenauer, and Michèle Sebag. 2012. April: Active preference learning-based reinforcement learning. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II 23*, pages 116–131. Springer.

[14] Jide Alaga and Jonas Schuett. 2023. Coordinated pausing: An evaluation-based coordination scheme for frontier ai developers. *arXiv preprint arXiv:2310.00374*.

[15] Guillaume Alain and Yoshua Bengio. 2017. Understanding intermediate layers using linear classifier probes. https://openreview.net/forum?id=ryF7rTqgl.

[16] Stefano V Albrecht and Subramanian Ramamoorthy. 2013. A game-theoretic model and best-response learning method for ad hoc coordination in multiagent systems. In *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*, pages 1155–1156.

[17] Gordon Willard Allport. 1955. *Becoming: Basic considerations for a psychology of personality*, volume 20. Yale University Press.

[18] David Alvarez Melis and Tommi Jaakkola. 2018. Towards robust interpretability with self-explaining neural networks. *Advances in Neural Information Processing Systems*, 31.

[19] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *AI Magazine*, 35(4):105–120.

[20] Dario Amodei, Paul Christiano, and Alex Ray. 2017. Learning from human preferences. https://openai.com/research/learning-from-human-preferences.

References

[21] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*.

[22] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. 2018. Towards better understanding of gradient-based attribution methods for deep neural networks. In *6th International Conference on Learning Representations (ICLR)*, 1711.06104, pages 0–0. Arxiv-Computer Science.

[23] Markus Anderljung, Joslyn Barnhart, Jade Leung, Anton Korinek, Cullen O'Keefe, Jess Whittlestone, Shahar Avin, Miles Brundage, Justin Bullock, Duncan Cass-Beggs, et al. 2023. Frontier ai regulation: Managing emerging risks to public safety. *arXiv preprint arXiv:2307.03718*.

[24] Michael Anderson, Susan Anderson, and Chris Armen. 2005. Towards machine ethics: Implementing two action-based ethical theories. In *Proceedings of the AAAI 2005 fall symposium on machine ethics*, pages 1–7.

[25] Michael Anderson and Susan Leigh Anderson. 2007. The status of machine ethics: a report from the aaai symposium. *Minds and Machines*, 17:1–10.

[26] Michael Anderson and Susan Leigh Anderson. 2011. *Machine ethics*. Cambridge University Press.

[27] Jacob Andreas. 2022. Language models as agent models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5769–5779.

[28] Anthropic. 2022. Softmax linear units. https://transformer-circuits.pub/2022/solu/index.html.

[29] Anthropic. 2023a. Anthropic's responsible scaling policy. https://www.anthropic.com/index/anthropics-responsible-scaling-policy.

[30] Anthropic. 2023b. Circuits updates - july 2023. https://transformer-circuits.pub/2023/july-update/index.html.

[31] Anthropic. 2023c. Model card and evaluations for claude models. https://www-files.anthropic.com/production/images/Model-Card-Claude-2.pdf.

[32] ARC Evals. 2023. Update on ARC's recent eval efforts. https://evals.alignment.org/blog/2023-03-18-update-on-recent-evals/.

[33] Sercan Ö Arik and Tomas Pfister. 2021. Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35(8), pages 6679–6687.

[34] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

[35] Konstantine Arkoudas, Selmer Bringsjord, and Paul Bello. 2005. Toward ethical robots via mechanized deontic logic. In *AAAI fall symposium on machine ethics*, pages 17–23. The AAAI Press Menlo Park, CA, USA.

[36] Stuart Armstrong. 2019. problems with ai debate. https://www.alignmentforum.org/posts/fNTCveSa4HvqvZR2F/problems-with-ai-debate.

[37] Stuart Armstrong, Nick Bostrom, and Carl Shulman. 2016. Racing to the precipice: a model of artificial intelligence development. *AI & society*, 31:201–206.

[38] Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2018. Linear algebraic structure of word senses, with applications to polysemy. *Transactions of the Association for Computational Linguistics*, 6:483–495.

[39] Saurabh Arora and Prashant Doshi. 2021. A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence*, 297:103500.

[40] Kenneth J Arrow. 2012. *Social choice and individual values*, volume 12. Yale university press.

[41] Asimov. 1942. Asimov's laws. https://webhome.auburn.edu/~vestmon/robotics.html.

[42] Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

[43] Karl Johan Åström and Richard M Murray. 2021. *Feedback systems: an introduction for scientists and engineers*. Princeton university press.

[44] Karl Johan Åström and Björn Wittenmark. 2008. *Adaptive control*. Courier Corporation.

[45] Atomium-EISMD. 2023. Ai4people. https://www.eismd.eu/ai4people.

[46] Alexandre Attia and Sharone Dayan. 2018. Global overview of imitation learning. *arXiv preprint arXiv:1801.06503*.

[47] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. The moral machine experiment. *Nature*, 563(7729):59–64.

[48] Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. 2023. A general theoretical paradigm to understand learning from human preferences. *arXiv preprint arXiv:2310.12036*.

[49] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

[50] Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. 2021. Recent advances in adversarial training for adversarial robustness. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4312–4321. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

[51] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

[52] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022b. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

[53] Michael Bain and Claude Sammut. 1995. A framework for behavioural cloning. In *Machine Intelligence 15*, pages 103–129.

[54] Andrea Bajcsy, Dylan P Losey, Marcia K O'Malley, and Anca D Dragan. 2018. Learning from physical human corrections, one feature at a time. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 141–149.

[55] Bowen Baker, Ilge Akkaya, Peter Zhokov, Joost Huizinga, Jie Tang, Adrien Ecoffet, Brandon Houghton, Raul Sampedro, and Jeff Clune. 2022. Video pretraining (vpt): Learning to act by watching unlabeled online videos. *Advances in Neural Information Processing Systems*, 35:24639–24654.

[56] Michiel Bakker, Martin Chadwick, Hannah Sheahan, Michael Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matt Botvinick, et al. 2022. Fine-tuning language models to find agreement among humans with diverse preferences. *Advances in Neural Information Processing Systems*, 35:38176–38189.

[57] Paul Bakker, Yasuo Kuniyoshi, et al. 1996. Robot see, robot do: An overview of robot imitation. In *AISB96 Workshop on Learning in Robots and Animals*, volume 5.

[58] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics.

[59] Yamini Bansal, Preetum Nakkiran, and Boaz Barak. 2021. Revisiting model stitching to compare neural representations. *Advances in Neural Information Processing Systems*, 34:225–236.

[60] Beth Barnes. 2020. debate-update-obfuscated-arguments-problem. https://www.alignmentforum.org/posts/PJLABqQ962hZEqhdB/debate-update-obfuscated-arguments-problem.

[61] Feras A Batarseh, Laura Freeman, and Chih-Hao Huang. 2021. A survey on artificial intelligence assurance. *Journal of Big Data*, 8(1):60.

[62] Sara Beery, Grant Van Horn, and Pietro Perona. 2018. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473.

[63] Mark Beliaev, Andy Shih, Stefano Ermon, Dorsa Sadigh, and Ramtin Pedarsani. 2022. Imitation learning by estimating expertise of demonstrators. In *International Conference on Machine Learning*, pages 1732–1748. PMLR.

[64] Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

[65] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. 2018. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*.

[66] Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*.

[67] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. 2009. *Robust optimization*, volume 28. Princeton university press.

[68] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

[69] Yoshua Bengio. 2023. How rogue ais may arise. `https://yoshuabengio.org/2023/05/22/how-rogue-ais-may-arise`.

[70] Yoshua Bengio, Stuart Russell, Elon Musk, and Future of Life Institute. 2023. Pause giant ai experiments: An open letter. `https://futureoflife.org/open-letter/pause-giant-ai-experiments`.

[71] Tsvi Benson-Tilsen and Nate Soares. 2016. Formalizing convergent instrumental goals. In *AAAI Workshop: AI, Ethics, and Society*.

[72] Gregory Benton, Wesley Maddox, Sanae Lotfi, and Andrew Gordon Gordon Wilson. 2021. Loss surface simplexes for mode connecting volumes and fast ensembling. In *International Conference on Machine Learning*, pages 769–779. PMLR.

[73] Hugo Berg, Siobhan Hall, Yash Bhalgat, Hannah Kirk, Aleksandar Shtedritski, and Max Bain. 2022. A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 806–822.

[74] Stuart Berg, Dominik Kutra, Thorben Kroeger, Christoph N Straehle, Bernhard X Kausler, Carsten Haubold, Martin Schiegg, Janez Ales, Thorsten Beier, Markus Rudy, et al. 2019. Ilastik: interactive machine learning for (bio) image analysis. *Nature methods*, 16(12):1226–1232.

[75] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2017. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409*.

[76] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2021. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44.

[77] Fiona Berreby, Gauvain Bourgne, and Jean-Gabriel Ganascia. 2017. A declarative modular framework for representing and applying ethical principles. In *16th Conference on Autonomous Agents and MultiAgent Systems*.

[78] Omar Besbes, Will Ma, and Omar Mouchtaki. 2022. Beyond IID: data-driven decision-making in heterogeneous environments. *Advances in Neural Information Processing Systems*, 35:23979–23991.

[79] Paul Christiano Beth Barnes. 2020. writeup-progress-on-ai-safety-via-debate-1. `https://www.alignmentforum.org/posts/Br4xDbYu4Frwrb64a/writeup-progress-on-ai-safety-via-debate-1`.

[80] Anand Bhattad, Min Jin Chong, Kaizhao Liang, Bo Li, and DA Forsyth. 2019. Unrestricted adversarial examples via semantic manipulation. In *International Conference on Learning Representations*.

[81] Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. 2023. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619(7970):533–538.

[82] Zhu Ming Bi, Chaomin Luo, Zhonghua Miao, Bing Zhang, WJ Zhang, and Lihui Wang. 2021. Safety assurance mechanisms of collaborative robotic systems in manufacturing. *Robotics and Computer-Integrated Manufacturing*, 67:102022.

[83] Richard Blumenthal and Josh Hawley. 2023. Bipartisan framework for u.s. ai act. `https://www.blumenthal.senate.gov/imo/media/doc/09072023bipartisanaiframework.pdf`.

[84] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in Neural Information Processing Systems*, 29.

[85] Eric Bonabeau. 2002. Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the national academy of sciences*, 99(suppl_3):7280–7287.

[86] Nick Bostrom. 2012. The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines*, 22:71–85.

[87] Nick Bostrom. 2013. Existential risk prevention as global priority. *Global Policy*, 4(1):15–31.

[88] Nick Bostrom and Milan M Cirkovic. 2011. *Global catastrophic risks*. Oxford University Press, USA.

[89] Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. 2021. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 141–159. IEEE.

[90] Samuel R Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamilė Lukošiūtė, Amanda Askell, Andy Jones, Anna Chen, et al. 2022. Measuring progress on scalable oversight for large language models. *arXiv preprint arXiv:2211.03540*.

[91] Hamed Bozorgi and Trung Dung Ngo. 2023. Beyond shared autonomy: Joint perception and action for human-in-the-loop mobile robot navigation systems. *Journal of Intelligent & Robotic Systems*, 109(1):20.

[92] Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

[93] Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D Procaccia. 2016. *Handbook of computational social choice*. Cambridge University Press.

[94] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, et al. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, page 2.

[95] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. 2016. Openai gym. *arXiv preprint arXiv:1606.01540*.

[96] Daniel Brown, Russell Coleman, Ravi Srinivasan, and Scott Niekum. 2020a. Safe imitation learning via fast bayesian reward inference from preferences. In *International Conference on Machine Learning*, pages 1165–1177. PMLR.

[97] Daniel S. Brown, Wonjoon Goo, Prabhat Nagarajan, and Scott Niekum. 2019. Extrapolating Beyond Suboptimal Demonstrations via Inverse Reinforcement Learning from Observations. In *International Conference on Machine Learning (ICML)*, pages 783–792.

[98] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020b. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

[99] Miles Brundage, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, Jingying Yang, Helen Toner, Ruth Fong, et al. 2020. Toward trustworthy ai development: mechanisms for supporting verifiable claims. *arXiv preprint arXiv:2004.07213*.

[100] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

[101] Alexander Bukharin, Yixiao Li, Pengcheng He, Weizhu Chen, and Tuo Zhao. 2023. Deep reinforcement learning from hierarchical weak preference feedback. *arXiv preprint arXiv:2309.02632*.

[102] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR.

References

[103] Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, et al. 2023. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv preprint arXiv:2312.09390*.

[104] Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2022. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations*.

[105] Lucian Buşoniu, Tim De Bruin, Domagoj Tolić, Jens Kober, and Ivana Palunko. 2018. Reinforcement learning for control: Performance, stability, and deep approximators. *Annual Reviews in Control*, 46:8–28.

[106] Serkan Cabi, Sergio Gómez Colmenarejo, Alexander Novikov, Ksenia Konyushkova, Scott E. Reed, Rae Jeong, Konrad Zolna, Yusuf Aytar, David Budden, Mel Vecerík, Oleg Sushkov, David Barker, Jonathan Scholz, Misha Denil, Nando de Freitas, and Ziyu Wang. 2020. Scaling data-driven robotics with reward sketching and batch reinforcement learning. In *Robotics: Science and Systems XVI, Virtual Event / Corvalis, Oregon, USA, July 12-16, 2020*.

[107] Ángel Alexander Cabrera, Erica Fu, Donald Bertucci, Kenneth Holstein, Ameet Talwalkar, Jason I Hong, and Adam Perer. 2023. Zeno: An interactive framework for behavioral evaluation of machine learning. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

[108] CAIS. 2023. Center for ai safety: Statement on ai risk. https://www.safe.ai/statement-on-ai-risk.

[109] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

[110] Rafael A Calvo, Dorian Peters, and Stephen Cave. 2020. Advancing impact assessment for intelligent systems. *Nature Machine Intelligence*, 2(2):89–91.

[111] Ella Cao and Eduardo Baptista. 2023. 'deepfake' scam in china fans worries over ai-driven fraud. *Reuters*.

[112] Nicholas Carlini, Milad Nasr, Christopher A Choquette-Choo, Matthew Jagielski, Irena Gao, Pang Wei W Koh, Daphne Ippolito, Florian Tramer, and Ludwig Schmidt. 2024. Are aligned neural networks adversarially aligned? *Advances in Neural Information Processing Systems*, 36.

[113] Joseph Carlsmith. 2022. Is power-seeking ai an existential risk? *arXiv preprint arXiv:2206.13353*.

[114] Tom Carlson and Yiannis Demiris. 2010. Increasing robotic wheelchair safety with collaborative control: Evidence from secondary task experiments. In *2010 IEEE International Conference on Robotics and Automation*, pages 5582–5587. IEEE.

[115] Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. 2019. Unlabeled data improves adversarial robustness. *Advances in Neural Information Processing Systems*, 32.

[116] Andrew Carr. 2023. Teaching large language models to zip their lips. https://gretel.ai/blog/teaching-large-language-models-to-zip-their-lips.

[117] Andres Carranza, Dhruv Pai, Rylan Schaeffer, Arnuv Tandon, and Sanmi Koyejo. 2023. Deceptive alignment monitoring.

[118] Micah Carroll, Alan Chan, Henry Ashton, and David Krueger. 2023. Characterizing manipulation from ai systems. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*.

[119] Micah D Carroll, Anca Dragan, Stuart Russell, and Dylan Hadfield-Menell. 2022. Estimating and penalizing induced preference shifts in recommender systems. In *International Conference on Machine Learning*, pages 2686–2708. PMLR.

[120] Shan Carter, Zan Armstrong, Ludwig Schubert, Ian Johnson, and Chris Olah. 2019. Activation atlas. *Distill*, 4(3):e15.

[121] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1721–1730.

[122] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. 2019. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832.

[123] Stephen Casper. 2023. Moving Forward: 11th post of The Engineer's Interpretability Sequence. https://www.alignmentforum.org/posts/L5Rua9aTndviy8dvc/eis-xi-moving-forward.

References

[124] Stephen Casper, Tong Bu, Yuxiao Li, Jiawei Li, Kevin Zhang, Kaivalya Hariharan, and Dylan Hadfield-Menell. 2023a. Red teaming deep neural networks with feature synthesis tools. In *Thirty-seventh Conference on Neural Information Processing Systems*.

[125] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Tong Wang, Samuel Marks, Charbel-Raphaël Segerie, Micah Carroll, Andi Peng, Phillip Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Biyik, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. 2023b. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research*. Survey Certification.

[126] Stephen Casper, Jason Lin, Joe Kwon, Gatlen Culp, and Dylan Hadfield-Menell. 2023c. Explore, establish, exploit: Red teaming language models from scratch. *arXiv preprint arXiv:2306.09442*.

[127] Stephen Casper, Max Nadeau, Dylan Hadfield-Menell, and Gabriel Kreiman. 2022. Robust feature-level adversaries are interpretability tools. *Advances in Neural Information Processing Systems*, 35:33093–33106.

[128] Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

[129] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. 2021. A survey on adversarial attacks and defences. *CAAI Transactions on Intelligence Technology*, 6(1):25–45.

[130] Souradip Chakraborty, Amrit Bedi, Alec Koppel, Huazheng Wang, Dinesh Manocha, Mengdi Wang, and Furong Huang. 2024. PARL: A unified framework for policy alignment in reinforcement learning. In *The Twelfth International Conference on Learning Representations*.

[131] Alan Chan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, Zhonghao He, Yawen Duan, Micah Carroll, et al. 2023. Harms from increasingly agentic algorithmic systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 651–666.

[132] Raja Chatila and John C Havens. 2019. The ieee global initiative on ethics of autonomous and intelligent systems. *Robotics and well-being*, pages 11–16.

[133] Pablo Chavez. 2023. An ai challenge: Balancing open and closed systems. https://cepa.org/artic le/an-ai-challenge-balancing-open-and-closed-systems.

[134] Hila Chefer, Shir Gur, and Lior Wolf. 2021. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 782–791.

[135] Canyu Chen and Kai Shu. 2024. Can LLM-generated misinformation be detected? In *The Twelfth International Conference on Learning Representations*.

[136] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

[137] Zhaoyu Chen, Bo Li, Shuang Wu, Kaixun Jiang, Shouhong Ding, and Wenqiang Zhang. 2024. Content-based unrestricted adversarial attack. *Advances in Neural Information Processing Systems*, 36.

[138] Minhao Cheng, Jinfeng Yi, Pin-Yu Chen, Huan Zhang, and Cho-Jui Hsieh. 2020. Seq2sick: Evaluating the robustness of sequence-to-sequence models with adversarial examples. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):3601–3608.

[139] Weiwei Cheng, Eyke Hüllermeier, and Krzysztof J Dembczynski. 2010a. Graded multilabel classification: The ordinal case. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 223–230.

[140] Weiwei Cheng, Eyke Hüllermeier, and Krzysztof J Dembczynski. 2010b. Label ranking methods based on the plackett-luce model. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 215–222.

[141] Weiwei Cheng, Michaël Rademaker, Bernard De Baets, and Eyke Hüllermeier. 2010c. Predicting partial orders: ranking with abstention. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010, Proceedings, Part I 21*, pages 215–230. Springer.

References

[142] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023).*

[143] Leshem Choshen, Lior Fox, Zohar Aizenbud, and Omri Abend. 2019. On the weaknesses of reinforcement learning for neural machine translation. *arXiv preprint arXiv:1907.01752.*

[144] Brian Christian. 2020. *The alignment problem: Machine learning and human values*. WW Norton & Company.

[145] Paul Christiano. 2019. What failure looks like. `https://www.alignmentforum.org/posts/HBx e6wdjxK239zajf/what-failure-looks-like`.

[146] Paul Christiano. 2022. Approval-directed agents. `https://www.alignmentforum.org/posts/7 Hr8t6xwuuxBTqADK/approval-directed-agents-1`.

[147] Paul Christiano. 2023. Thoughts on the impact of rlhf research. `https://www.lesswrong.com/po sts/vwu4kegAEZTBtpT6p/thoughts-on-the-impact-of-rlhf-research`.

[148] Paul Christiano, Buck Shlegeris, and Dario Amodei. 2018. Supervising strong learners by amplifying weak experts. *arXiv preprint arXiv:1810.08575.*

[149] Paul Christiano, Mark Xu, and Ajeya Cotra. 2021. Arc's first technical report: Eliciting latent knowledge. `https://www.alignmentforum.org/posts/qHCDysDnvhteW7kRd/arc-s-first-techn ical-report-eliciting-latent-knowledge`.

[150] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30.

[151] Phillip JK Christoffersen, Andreas A Haupt, and Dylan Hadfield-Menell. 2023. Get it in writing: Formal contracts mitigate social dilemmas in multi-agent rl. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pages 448–456.

[152] Bilal Chughtai, Lawrence Chan, and Neel Nanda. 2023. A toy model of universality: Reverse engineering how networks learn group operations. *arXiv preprint arXiv:2302.03025.*

[153] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does bert look at? an analysis of bert's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286.

[154] Ruby RobertM GPT-4 Claude. 2023. New lw feature debates. `https://www.lesswrong.com/post s/kXiAGRWFquXFMi68Y/new-lw-feature-debates`.

[155] Code Bullet. 2019. Simulator with bugs. `https://www.youtube.com/watch?v=K-wIZuAA3EY`.

[156] Collective Intelligence Project. 2023. Introducing the collective intelligence project. `https://cip.or g/whitepaper`.

[157] Vincent Conitzer, Walter Sinnott-Armstrong, Jana Schaich Borg, Yuan Deng, and Max Kramer. 2017. Moral decision making frameworks for artificial intelligence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

[158] Arthur Conmy, Augustine N Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability. *arXiv preprint arXiv:2304.14997.*

[159] Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. 2009. L2 regularization for learning kernels. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence, UAI 2009*, pages 109–116. AUAI Press.

[160] Ajeya Cotra. 2018. Iterated distillation and amplification.

[161] Ajeya Cotra. 2021. The case for aligning narrowly superhuman models. `https://www.alignmentf orum.org/posts/PZtsoaoSLpKjjbMqM/the-case-for-aligning-narrowly-superhuma n-models`.

[162] Ajeya Cotra. 2022. Without specific countermeasures, the easiest path to transformative AI likely leads to AI takeover - AI Alignment Forum. `https://www.alignmentforum.org/posts/pRkFkzwKZ2zfa3R 6H/without-specific-countermeasures-the-easiest-path-to`.

References

[163] Andrew Critch and David Krueger. 2020. Ai research considerations for human existential safety (arches). *arXiv preprint arXiv:2006.04948*.

[164] Andrew Critch and Stuart Russell. 2023. Tasra: A taxonomy and analysis of societal-scale risks from ai. *arXiv preprint arXiv:2306.06924*.

[165] Diogo Cruz, José Aleixo Cruz, and Henrique Lopes Cardoso. 2019. Reinforcement learning in multi-agent games: Open ai gym diplomacy environment. In *Progress in Artificial Intelligence: 19th EPIA Conference on Artificial Intelligence, EPIA 2019, Vila Real, Portugal, September 3–6, 2019, Proceedings, Part I 19*, pages 49–60. Springer.

[166] Brandon Cui, Hengyuan Hu, Luis Pineda, and Jakob Foerster. 2021. K-level reasoning for zero-shot coordination in hanabi. *Advances in Neural Information Processing Systems*, 34:8215–8228.

[167] Allan Dafoe, Yoram Bachrach, Gillian Hadfield, Eric Horvitz, Kate Larson, and Thore Graepel. 2021. Cooperative ai: machines must learn to find common ground. *Nature*, 593(7857):33–36.

[168] Allan Dafoe, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R McKee, Joel Z Leibo, Kate Larson, and Thore Graepel. 2020. Open problems in cooperative ai. *arXiv preprint arXiv:2012.08630*.

[169] Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502.

[170] Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. 2023. Why can gpt learn in-context? language models implicitly perform gradient descent as meta-optimizers. In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*.

[171] Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. 2024. Safe rlhf: Safe reinforcement learning from human feedback. In *The Twelfth International Conference on Learning Representations*.

[172] Brian d'Alessandro, Cathy O'Neil, and Tom LaGatta. 2017. Conscientious classification: A data scientist's guide to discrimination-aware classification. *Big data*, 5(2):120–134.

[173] Corentin Dancette, Remi Cadene, Damien Teney, and Matthieu Cord. 2021. Beyond question-based biases: Assessing multimodal shortcut learning in visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1574–1583.

[174] Richard Danzig. 2012. Aum shinrikyo: insights into how terrorists develop biological and chemical weapons. *Studies in Conflict & Terrorism*.

[175] Guy Dar, Mor Geva, Ankit Gupta, and Jonathan Berant. 2023. Analyzing transformers in embedding space. In *Annual Meeting of the Association for Computational Linguistics*.

[176] Sudeep Dasari, Abhinav Gupta, and Vikash Kumar. 2023. Learning Dexterous Manipulation from Exemplar Object Trajectories and Pre-Grasps. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE.

[177] Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations*.

[178] Pim De Haan, Dinesh Jayaraman, and Sergey Levine. 2019. Causal confusion in imitation learning. *Advances in Neural Information Processing Systems*, 32.

[179] Scott De Marchi and Scott E Page. 2014. Agent-based models. *Annual Review of political science*, 17:1–20.

[180] DeepMind. 2018. Building safe artificial intelligence: specification, robustness, and assurance. https://deepmindsafetyresearch.medium.com/building-safe-artificial-intelligence-52f5f75058f1.

[181] DeepMind. 2020. goal misgeneralization. https://docs.google.com/spreadsheets/u/1/d/e/2PACX-1vTo3RkXUAigb25nP7gjpcHriR6XdzA_L5loOcVFj_u7cRAZghWrYKH2L2nU4TA_Vr9KzBX5Bjpz9G_l/pubhtml?pli=1.

[182] Jonas Degrave, Federico Felici, Jonas Buchli, Michael Neunert, Brendan Tracey, Francesco Carpanese, Timo Ewalds, Roland Hafner, Abbas Abdolmaleki, Diego de Las Casas, et al. 2022. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*, 602(7897):414–419.

[183] Gelei Deng, Yi Liu, Yuekang Li, Kailong Wang, Ying Zhang, Zefeng Li, Haoyu Wang, Tianwei Zhang, and Yang Liu. 2023. Jailbreaker: Automated jailbreak across multiple large language model chatbots. *arXiv preprint arXiv:2307.08715*.

[184] Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric Xing, and Zhiting Hu. 2022. Rlprompt: Optimizing discrete text prompts with reinforcement learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3369–3391.

[185] Louise Dennis, Michael Fisher, Marija Slavkovik, and Matt Webster. 2016. Formal verification of ethical choices in autonomous systems. *Robotics and Autonomous Systems*, 77:1–14.

[186] Michael Dennis, Natasha Jaques, Eugene Vinitsky, Alexandre Bayen, Stuart Russell, Andrew Critch, and Sergey Levine. 2020. Emergent complexity and zero-shot transfer via unsupervised environment design. *Advances in Neural Information Processing Systems*, 33:13049–13061.

[187] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2020. Eraser: A benchmark to evaluate rationalized nlp models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458.

[188] Bhuwan Dhingra, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das, and William Cohen. 2019. Handling divergent reference texts when evaluating table-to-text generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4884–4895.

[189] Lauro Langosco Di Langosco, Jack Koch, Lee D Sharkey, Jacob Pfau, and David Krueger. 2022. Goal misgeneralization in deep reinforcement learning. In *International Conference on Machine Learning*, pages 12004–12019. PMLR.

[190] Thomas G Dietterich. 2017. Steps toward robust artificial intelligence. *AI Magazine*, 38(3):3–24.

[191] Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, KaShun SHUM, and Tong Zhang. 2023. RAFT: Reward ranked finetuning for generative foundation model alignment. *Transactions on Machine Learning Research*.

[192] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.

[193] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

[194] Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred Hamprecht. 2018. Essentially no barriers in neural network energy landscape. In *International conference on machine learning*, pages 1309–1318. PMLR.

[195] Mengnan Du, Ninghao Liu, and Xia Hu. 2019. Techniques for interpretable machine learning. *Communications of the ACM*, 63(1):68–77.

[196] Yali Du. 2023. Cooperative multi-agent learning in a complex world: challenges and solutions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37(13), pages 15436–15436.

[197] Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

[198] Veljko Dubljevic. 2020. Toward implementing the agent-deed-consequence model of moral judgment in autonomous vehicles. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 243–243.

[199] John C Duchi, Peter W Glynn, and Hongseok Namkoong. 2021. Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 46(3):946–969.

[200] John C Duchi, Lester W Mackey, and Michael I Jordan. 2010. On the consistency of ranking algorithms. In *ICML*, pages 327–334.

[201] Nadir Durrani, Hassan Sajjad, Fahim Dalvi, and Yonatan Belinkov. 2020. Analyzing individual neurons in pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4865–4880.

[202] Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. Hotflip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 31–36.

[203] Mark Edmonds, Feng Gao, Xu Xie, Hangxin Liu, Siyuan Qi, Yixin Zhu, Brandon Rothrock, and Song-Chun Zhu. 2017. Feeling the force: Integrating force and pose for fluent discovery through imitation learning to open medicine bottles. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3530–3537. IEEE.

[204] Ronen Eldan and Mark Russinovich. 2023. Who's harry potter? approximate unlearning in llms. *arXiv preprint arXiv:2310.02238*.

[205] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. 2022. Toy models of superposition. *arXiv preprint arXiv:2209.10652*.

[206] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1.

[207] Daniel C Elton. 2020. Self-explaining ai as an alternative to interpretable ai. In *Artificial General Intelligence: 13th International Conference, AGI 2020, St. Petersburg, Russia, September 16–19, 2020, Proceedings 13*, pages 95–106. Springer.

[208] Denis Emelin, Ronan Le Bras, Jena D Hwang, Maxwell Forbes, and Yejin Choi. 2021. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 698–718.

[209] Logan Engstrom, Andrew Ilyas, Hadi Salman, Shibani Santurkar, and Dimitris Tsipras. 2019a. Robustness (python library). https://github.com/MadryLab/robustness.

[210] Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Jacob Steinhardt, and Aleksander Madry. 2020. Identifying statistical bias in dataset replication. In *International Conference on Machine Learning*, pages 2922–2932. PMLR.

[211] Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. 2019b. Exploring the landscape of spatial robustness. In *International conference on machine learning*, pages 1802–1811. PMLR.

[212] Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2009. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1.

[213] Eva Erman and Markus Furendal. 2022. Artificial intelligence and the political legitimacy of global governance. *Political Studies*, page 00323217221126665.

[214] Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Zohreh Shams, Frederic Precioso, Stefano Melacci, Adrian Weller, et al. 2022. Concept embedding models: Beyond the accuracy-explainability trade-off. *Advances in Neural Information Processing Systems*, 35:21400–21413.

[215] European Parliament. 2023. Eu ai act: first regulation on artificial intelligence. https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence.

[216] Evan Hubinger. 2023. Bing chat is blatantly, aggressively misaligned. https://www.lesswrong.com/posts/jtoPawEhLNXNxvgTT/bing-chat-is-blatantly-aggressively-misaligned.

[217] Tom Everitt and Marcus Hutter. 2016. Avoiding wireheading with value reinforcement learning. In *Artificial General Intelligence: 9th International Conference, AGI 2016, New York, NY, USA, July 16-19, 2016, Proceedings 9*, pages 12–22. Springer.

[218] Tom Everitt, Marcus Hutter, Ramana Kumar, and Victoria Krakovna. 2021. Reward tampering problems and solutions in reinforcement learning: A causal influence diagram perspective. *Synthese*, 198(Suppl 27):6435–6467.

[219] Tom Everitt, Victoria Krakovna, Laurent Orseau, and Shane Legg. 2017. Reinforcement learning with a corrupted reward channel. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 4705–4713.

[220] Tom Everitt, Gary Lea, and Marcus Hutter. 2018. Agi safety literature review. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 5441–5449. International Joint Conferences on Artificial Intelligence Organization.

[221] Benjamin Eysenbach, Shixiang Gu, Julian Ibarz, and Sergey Levine. 2018. Leave no trace: Learning to reset for safe and autonomous reinforcement learning. In *International Conference on Learning Representations*.

[222] Daniel Fabian. 2023. Google's ai red team: the ethical hackers making ai safer. `https://blog.googl e/technology/safety-security/googles-ai-red-team-the-ethical-hackers-mak ing-ai-safer`.

[223] Jerry Alan Fails and Dan R Olsen Jr. 2003. Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces*, pages 39–45.

[224] Diplomacy Team FAIR, Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, et al. 2022. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074.

[225] Tobias Falke, Leonardo FR Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 2214–2220.

[226] Feng-Lei Fan, Jinjun Xiong, Mengzhou Li, and Ge Wang. 2021. On interpretability of artificial neural networks: A survey. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 5(6):741–760.

[227] Bin Fang, Shidong Jia, Di Guo, Muhua Xu, Shuhuan Wen, and Fuchun Sun. 2019. Survey of imitation learning for robotic manipulation. *International Journal of Intelligent Robotics and Applications*, 3:362–369.

[228] Alhussein Fawzi, Matej Balog, Aja Huang, Thomas Hubert, Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Francisco J R Ruiz, Julian Schrittwieser, Grzegorz Swirszcz, et al. 2022. Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature*, 610(7930):47–53.

[229] Patrick Fernandes, Aman Madaan, Emmy Liu, António Farinhas, Pedro Henrique Martins, Amanda Bertsch, José GC de Souza, Shuyan Zhou, Tongshuang Wu, Graham Neubig, et al. 2023. Bridging the gap: A survey on integrating (human) feedback for natural language generation. *arXiv preprint arXiv:2305.00955*.

[230] Pedro M Fernandes, Francisco C Santos, and Manuel Lopes. 2020. Adoption dynamics and societal impact of ai systems in complex networks. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 258–264.

[231] Arnaud Fickinger, Simon Zhuang, Dylan Hadfield-Menell, and Stuart Russell. 2020. Multi-principal assistance games. *arXiv preprint arXiv:2007.09540*.

[232] Tanner Fiez, Benjamin Chasnov, and Lillian Ratliff. 2020. Implicit learning dynamics in stackelberg games: Equilibria characterization, convergence analysis, and empirical study. In *International Conference on Machine Learning*, pages 3133–3144. PMLR.

[233] Jaime F Fisac, Monica A Gates, Jessica B Hamrick, Chang Liu, Dylan Hadfield-Menell, Malayandi Palaniappan, Dhruv Malik, S Shankar Sastry, Thomas L Griffiths, and Anca D Dragan. 2020. Pragmatic-pedagogic value alignment. In *Robotics Research: The 18th International Symposium ISRR*, pages 49–57. Springer.

[234] Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, et al. 2021. An ethical framework for a good ai society: Opportunities, risks, principles, and recommendations. *Ethics, governance, and policies in artificial intelligence*, pages 19–39.

[235] Jakob Foerster, Ioannis Alexandros Assael, Nando De Freitas, and Shimon Whiteson. 2016. Learning to communicate with deep multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 29.

[236] Jakob N Foerster, Justin Gilmer, Jascha Sohl-Dickstein, Jan Chorowski, and David Sussillo. 2017. Input switched affine networks: An rnn architecture designed for interpretability. In *International conference on machine learning*, pages 1136–1145. PMLR.

[237] Ruth Fong and Andrea Vedaldi. 2018. Net2vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8730–8738.

[238] Maxwell Forbes, Jena D Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670.

[239] Jonathan Frankle, Gintare Karolina Dziugaite, Daniel Roy, and Michael Carbin. 2020. Linear mode connectivity and the lottery ticket hypothesis. In *International Conference on Machine Learning*, pages 3259–3269. PMLR.

[240] Daniel Freeman, David Ha, and Luke Metz. 2019. Learning to predict without looking ahead: World models without forward prediction. *Advances in Neural Information Processing Systems*, 32.

[241] Justin Fu, Anoop Korattikara, Sergey Levine, and Sergio Guadarrama. 2019. From language to goals: Inverse reinforcement learning for vision-based instruction following. In *International Conference on Learning Representations*.

[242] Justin Fu, Katie Luo, and Sergey Levine. 2018a. Learning robust rewards with adverserial inverse reinforcement learning. In *International Conference on Learning Representations*.

[243] Justin Fu, Avi Singh, Dibya Ghosh, Larry Yang, and Sergey Levine. 2018b. Variational inverse control with events: A general framework for data-driven reward definition. *Advances in Neural Information Processing Systems*, 31.

[244] Jason Furman and Robert Seamans. 2019. Ai and the economy. *Innovation policy and the economy*, 19(1):161–191.

[245] Johannes Fürnkranz and Eyke Hüllermeier. 2003. Pairwise preference learning and ranking. In *European conference on machine learning*, pages 145–156. Springer.

[246] Johannes Fürnkranz and Eyke Hüllermeier. 2010. *Preference Learning*. Springer Science & Business Media.

[247] G20. 2019. G20 ai principles. https://www.mofa.go.jp/policy/economy/g20_summit/osaka19/pdf/documents/en/annex_08.pdf.

[248] Iason Gabriel. 2020. Artificial intelligence, values, and alignment. *Minds and Machines*, 30(3):411–437.

[249] Iason Gabriel and Vafa Ghazavi. 2021. The challenge of value alignment: From fairer algorithms to ai safety. *arXiv preprint arXiv:2101.06060*.

[250] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. 2020. Large-scale adversarial training for vision-and-language representation learning. *Advances in Neural Information Processing Systems*, 33:6616–6628.

[251] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*.

[252] Leo Gao, John Schulman, and Jacob Hilton. 2023. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pages 10835–10866. PMLR.

[253] Javier Garcıa and Fernando Fernández. 2015. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480.

[254] Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P Vetrov, and Andrew G Wilson. 2018. Loss surfaces, mode connectivity, and fast ensembling of dnns. *Advances in Neural Information Processing Systems*, 31.

[255] Scott Garrabrant, Tsvi Benson-Tilsen, Andrew Critch, Nate Soares, and Jessica Taylor. 2016. Logical induction. *arXiv preprint arXiv:1609.03543*.

[256] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369.

[257] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. 2019. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*.

[258] Mor Geva, Avi Caciularu, Guy Dar, Paul Roit, Shoval Sadde, Micah Shlain, Bar Tamir, and Yoav Goldberg. 2022. Lm-debugger: An interactive tool for inspection and intervention in transformer-based language models. In *Proceedings of the The 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 12–21.

References

[259] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495.

[260] Seyed Kamyar Seyed Ghasemipour, Richard Zemel, and Shixiang Gu. 2020. A divergence minimization perspective on imitation learning methods. In *Conference on Robot Learning*, pages 1259–1277. PMLR.

[261] Justin Gilmer, Nicolas Ford, Nicholas Carlini, and Ekin Cubuk. 2019. Adversarial examples are a natural consequence of test error in noise. In *International Conference on Machine Learning*, pages 2280–2289. PMLR.

[262] Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. 2022. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*.

[263] Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. 2021. Multimodal neurons in artificial neural networks. *Distill*, 6(3):e30.

[264] Josh A Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. 2023. Generative language models and automated influence operations: Emerging threats and potential mitigations. *arXiv preprint arXiv:2301.04246*.

[265] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

[266] Charles AE Goodhart and CAE Goodhart. 1984. *Problems of monetary management: the UK experience*. Springer.

[267] Government of the United Kingdom. 2021. The roadmap to an effective ai assurance ecosystem - extended version. `https://www.gov.uk/government/publications/the-roadmap-to-an-effective-ai-assurance-ecosystem/the-roadmap-to-an-effective-ai-assurance-ecosystem-extended-version`.

[268] Government of the United Kingdom. 2023. Frontier ai: capabilities and risks – discussion paper. `https://www.gov.uk/government/publications/frontier-ai-capabilities-and-risks-discussion-paper`.

[269] Prasoon Goyal, Scott Niekum, and Raymond J Mooney. 2019. Using natural language for reward shaping in reinforcement learning. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 2385–2391.

[270] Nico Grant and Karen Weise. 2023. In ai race, microsoft and google choose speed over caution. *The New York Times*.

[271] Ryan Greenblatt, Buck Shlegeris, Kshitij Sachan, and Fabien Roger. 2023. Ai control: Improving safety despite intentional subversion. *arXiv preprint arXiv:2312.06942*.

[272] Shane Griffith, Kaushik Subramanian, Jonathan Scholz, Charles L Isbell, and Andrea L Thomaz. 2013. Policy shaping: Integrating human feedback with reinforcement learning. *Advances in Neural Information Processing Systems*, 26.

[273] Sven Gronauer and Klaus Diepold. 2022. Multi-agent deep reinforcement learning: a survey. *Artificial Intelligence Review*, pages 1–49.

[274] Roger Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, et al. 2023. Studying large language model generalization with influence functions. *arXiv preprint arXiv:2308.03296*.

[275] Luke Guerdan, Amanda Coston, Zhiwei Steven Wu, and Kenneth Holstein. 2023. Ground (less) truth: A causal framework for proxy labels in human-algorithm decision-making. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 688–704.

[276] Carlos Guestrin, Daphne Koller, and Ronald Parr. 2001. Multiagent planning with factored mdps. *Advances in Neural Information Processing Systems*, 14.

[277] Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, et al. 2023. Reinforced self-training (rest) for language modeling. *arXiv preprint arXiv:2308.08998*.

[278] Wes Gurnee and Max Tegmark. 2023. Language models represent space and time.

References

[279] António Guterres. 2023. Secretary-general's remarks to the security council on artificial intelligence. https://www.un.org/sg/en/content/sg/speeches/2023-07-18/secretary-generals-remarks-the-security-council-artificial-intelligence.

[280] Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell. 2017a. The off-switch game. In *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*.

[281] Dylan Hadfield-Menell and Gillian K Hadfield. 2019. Incomplete contracting and ai alignment. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 417–422.

[282] Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, Stuart J Russell, and Anca Dragan. 2017b. Inverse reward design. *Advances in Neural Information Processing Systems*, 30.

[283] Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. 2016. Cooperative inverse reinforcement learning. *Advances in Neural Information Processing Systems*, 29.

[284] Thilo Hagendorff. 2020. The ethics of ai ethics: An evaluation of guidelines. *Minds and Machines*, 30(1):99–120.

[285] Thilo Hagendorff. 2022. A virtue-based framework to support putting ai ethics into practice. *Philosophy & Technology*, 35(3):55.

[286] Michael Hanna, Ollie Liu, and Alexandre Variengien. 2024. How does gpt-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. *Advances in Neural Information Processing Systems*, 36.

[287] Yaru Hao, Li Dong, Furu Wei, and Ke Xu. 2021. Self-attention attribution: Interpreting information interactions inside transformer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12963–12971.

[288] Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326.

[289] Trevor Hastie, Robert Tibshirani, Jerome Friedman, Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. Overview of supervised learning. *The elements of statistical learning: Data mining, inference, and prediction*, pages 9–41.

[290] Jerry Zhi-Yang He and Anca D. Dragan. 2021. Assisted robust reward design. In *Conference on Robot Learning, 8-11 November 2021, London, UK*, volume 164 of *Proceedings of Machine Learning Research*, pages 1234–1246. PMLR.

[291] Stefan Heimersheim and Janiak Jett. 2023. A circuit for Python docstrings in a 4-layer attention-only transformer. https://www.alignmentforum.org/posts/u6KXXmKFbXfWzoAXn/a-circuit-for-python-docstrings-in-a-4-layer-attention-only.

[292] Joey Hejna, Rafael Rafailov, Harshit Sikchi, Chelsea Finn, Scott Niekum, W. Bradley Knox, and Dorsa Sadigh. 2024. Contrastive preference learning: Learning from human feedback without reinforcement learning. In *The Twelfth International Conference on Learning Representations*.

[293] Donald Joseph Hejna III and Dorsa Sadigh. 2022. Few-Shot Preference Learning for Human-in-the-Loop RL. In *Conference on Robot Learning (CoRL)*, pages 2014–2025.

[294] Dan Hendrycks. 2022. Pragmatic ai safety. https://www.alignmentforum.org/s/FaEBwhhe3otzYKGQt.

[295] Dan Hendrycks. 2023. Natural selection favors ais over humans.

[296] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. 2021a. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349.

[297] Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2020. Aligning ai with shared human values. In *International Conference on Learning Representations*.

[298] Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. 2021b. Unsolved problems in ml safety. *arXiv preprint arXiv:2109.13916*.

References

[299] Dan Hendrycks and Thomas Dietterich. 2018. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*.

[300] Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

[301] Dan Hendrycks and Mantas Mazeika. 2022. X-risk analysis for ai research. *arXiv preprint arXiv:2206.05862*.

[302] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. 2019. Using self-supervised learning can improve model robustness and uncertainty. *Advances in Neural Information Processing Systems*, 32.

[303] Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. 2023. An overview of catastrophic ai risks. *arXiv preprint arXiv:2306.12001*.

[304] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. 2021c. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271.

[305] Jonathan Ho and Stefano Ermon. 2016. Generative adversarial imitation learning. *Advances in Neural Information Processing Systems*, 29.

[306] Lewis Ho, Joslyn Barnhart, Robert Trager, Yoshua Bengio, Miles Brundage, Allison Carnegie, Rumman Chowdhury, Allan Dafoe, Gillian Hadfield, Margaret Levi, et al. 2023. International institutions for advanced ai. *arXiv preprint arXiv:2307.04699*.

[307] Marius Hobbhahn. 2022. Eliciting latent knowledge (elk) - distillation/summary. https://www.alignm entforum.org/posts/rxoBY9CMkqDsHt25t/eliciting-latent-knowledge-elk-disti llation-summary.

[308] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. An empirical analysis of compute-optimal large language model training. *Advances in Neural Information Processing Systems*, 35:30016–30030.

[309] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

[310] Andreas Holzinger. 2016. Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics*, 3(2):119–131.

[311] Andreas Holzinger, Chris Biemann, Constantinos S Pattichis, and Douglas B Kell. 2017. What do we need to build explainable ai systems for the medical domain? *arXiv preprint arXiv:1712.09923*.

[312] Joey Hong, Kush Bhatia, and Anca Dragan. 2022. On the sensitivity of reward inference to misspecified human models. In *The Eleventh International Conference on Learning Representations*.

[313] Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. Q2:: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7856–7870.

[314] Jeremy Howard. 2023. Ai safety and the age of dislightenment. https://www.fast.ai/posts/20 23-11-07-dislightenment.html.

[315] Hengyuan Hu, Adam Lerer, Brandon Cui, Luis Pineda, Noam Brown, and Jakob Foerster. 2021. Off-belief learning. In *International Conference on Machine Learning*, pages 4369–4379. PMLR.

[316] Hengyuan Hu, Adam Lerer, Alex Peysakhovich, and Jakob Foerster. 2020. "other-play" for zero-shot coordination. In *International Conference on Machine Learning*, pages 4399–4410. PMLR.

[317] Yingdong Hu, Renhao Wang, Li Erran Li, and Yang Gao. 2023. For Pre-Trained Vision Models in Motor Control, Not All Policy Learning Methods are Created Equal. In *International Conference on Machine Learning (ICML)*, pages 13628–13651.

[318] Sandy H. Huang, Nicolas Papernot, Ian J. Goodfellow, Yan Duan, and Pieter Abbeel. 2017. Adversarial attacks on neural network policies. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net.

[319] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. 2023. Inner monologue: Embodied reasoning through planning with language models. In *Conference on Robot Learning*, pages 1769–1782. PMLR.

[320] Yangsibo Huang, Samyak Gupta, Mengzhou Xia, Kai Li, and Danqi Chen. 2024. Catastrophic jailbreak of open-source LLMs via exploiting generation. In *The Twelfth International Conference on Learning Representations*.

[321] Evan Hubinger. 2020. An overview of 11 proposals for building safe advanced ai. *arXiv preprint arXiv:2012.07532*.

[322] Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M Ziegler, Tim Maxwell, Newton Cheng, et al. 2024. Sleeper agents: Training deceptive llms that persist through safety training. *arXiv preprint arXiv:2401.05566*.

[323] Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. 2019a. Deceptive alignment. https://www.alignmentforum.org/posts/zthDPAjh9w6Ytbeks/deceptive-alignment.

[324] Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. 2019b. The inner alignment problem. https://www.alignmentforum.org/posts/pL56xPoniLvtMDQ4J/the-inner-alignment-problem.

[325] Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. 2019c. Risks from learned optimization in advanced machine learning systems. *arXiv preprint arXiv:1906.01820*.

[326] Drew Arad Hudson and Christopher D. Manning. 2018. Compositional attention networks for machine reasoning. In *International Conference on Learning Representations*.

[327] Eyke Hüllermeier, Johannes Fürnkranz, Weiwei Cheng, and Klaus Brinker. 2008. Label ranking by learning pairwise preferences. *Artificial Intelligence*, 172(16-17):1897–1916.

[328] Ahmed Hussein, Mohamed Medhat Gaber, Eyad Elyan, and Chrisina Jayne. 2017. Imitation learning: A survey of learning methods. *ACM Computing Surveys (CSUR)*, 50(2):1–35.

[329] Borja Ibarz, Jan Leike, Tobias Pohlen, Geoffrey Irving, Shane Legg, and Dario Amodei. 2018. Reward learning from human preferences and demonstrations in atari. *Advances in Neural Information Processing Systems*, 31.

[330] BlueDot Impact. 2023. Alignment course. https://course.aisafetyfundamentals.com/alignment.

[331] Geoffrey Irving, Paul Christiano, and Dario Amodei. 2018. Ai safety via debate. *arXiv preprint arXiv:1805.00899*.

[332] Charles Isbell, Christian R Shelton, Michael Kearns, Satinder Singh, and Peter Stone. 2001. A social reinforcement learning agent. In *Proceedings of the fifth international conference on Autonomous agents*, pages 377–384.

[333] Jacob Steinhardt. 2023. Emergent deception and emergent optimization. https://bounded-regret.ghost.io/emergent-deception-optimization.

[334] Max Jaderberg, Wojciech M Czarnecki, Iain Dunning, Luke Marris, Guy Lever, Antonio Garcia Castaneda, Charles Beattie, Neil C Rabinowitz, Ari S Morcos, Avraham Ruderman, et al. 2019. Human-level performance in 3d multiplayer games with population-based reinforcement learning. *Science*, 364(6443):859–865.

[335] Ashesh Jain, Brian Wojcik, Thorsten Joachims, and Ashutosh Saxena. 2013. Learning trajectory preferences for manipulators via iterative improvement. *Advances in Neural Information Processing Systems*, 26.

[336] Hong Jun Jeon, Smitha Milli, and Anca Dragan. 2020. Reward-rational (implicit) choice: A unifying formalism for reward learning. *Advances in Neural Information Processing Systems*, 33:4415–4426.

[337] Jiaming Ji, Boyuan Chen, Hantao Lou, Donghai Hong, Borong Zhang, Xuehai Pan, Juntao Dai, and Yaodong Yang. 2024a. Aligner: Achieving efficient alignment through weak-to-strong correction. *arXiv preprint arXiv:2402.02416*.

[338] Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. 2024b. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*, 36.

[339] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys (CSUR)*, 55(12):1–38.

[340] Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031.

[341] Minqi Jiang, Michael Dennis, Jack Parker-Holder, Jakob Foerster, Edward Grefenstette, and Tim Rocktäschel. 2021. Replay-guided adversarial environment design. *Advances in Neural Information Processing Systems*, 34:1884–1897.

[342] Zhijing Jin, Sydney Levine, Fernando Gonzalez Adauto, Ojasv Kamal, Maarten Sap, Mrinmaya Sachan, Rada Mihalcea, Josh Tenenbaum, and Bernhard Schölkopf. 2022. When to make exceptions: Exploring language models as accounts of human moral judgment. *Advances in Neural Information Processing Systems*, 35:28458–28473.

[343] Jonas DeGrave. 2022. Building a virtual machine inside chatgpt. https://www.engraved.blog/building-a-virtual-machine-inside.

[344] Erik Jones, Anca D. Dragan, Aditi Raghunathan, and Jacob Steinhardt. 2023. Automatically auditing large language models via discrete optimization. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 15307–15329. PMLR.

[345] Michael I Jordan and Tom M Mitchell. 2015. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260.

[346] Jeevesh Juneja, Rachit Bansal, Kyunghyun Cho, João Sedoc, and Naomi Saphra. 2022. Linear connectivity reveals generalization strategies. In *The Eleventh International Conference on Learning Representations*.

[347] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. 1996. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285.

[348] Dimitris Kalimeris, Smriti Bhagat, Shankar Kalyanaraman, and Udi Weinsberg. 2021. Preference amplification in recommender systems. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 805–815.

[349] Josh Kalin, Matthew Ciolino, David Noever, and Gerry Dozier. 2020. Black box to white box: Discover model characteristics based on strategic probing. In *2020 Third International Conference on Artificial Intelligence for Industries (AI4I)*, pages 60–63. IEEE.

[350] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

[351] Andrej Karpathy, Justin Johnson, and Li Fei-Fei. 2015. Visualizing and understanding recurrent networks. *arXiv preprint arXiv:1506.02078*.

[352] Michael Kearns and Aaron Roth. 2019. *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press.

[353] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

[354] Zachary Kenton, Tom Everitt, Laura Weidinger, Iason Gabriel, Vladimir Mikulik, and Geoffrey Irving. 2021. Alignment of language agents. *arXiv preprint arXiv:2103.14659*.

[355] Zachary Kenton, Rohin Shah, David Lindner, Vikrant Varma, Victoria Krakovna, Mary Phuong, Ramana Kumar, and Elliot Catt. 2022. Threat model literature review. https://www.alignmentforum.org/posts/wnnkD6P2k2TfHnNmt/threat-model-literature-review.

[356] B. Kenward and T. R. Sinclair. 2021. Machine morality, moral progress, and the looming environmental disaster. *Cognitive Computation and Systems*, 3:83–90.

[357] Cameron F Kerry, Joshua P Meltzer, Andrea Renda, Alex Engler, and Rosanna Fanni. 2021. Strengthening international cooperation on ai, progress report. https://www.brookings.edu/articles/strengthening-international-cooperation-on-ai.

[358] Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R Bowman, Tim Rocktäschel, and Ethan Perez. 2024. Debating with more persuasive llms leads to more truthful answers. *arXiv preprint arXiv:2402.06782*.

[359] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR.

[360] Changyeon Kim, Jongjin Park, Jinwoo Shin, Honglak Lee, Pieter Abbeel, and Kimin Lee. 2023. Preference Transformer: Modeling Human Preferences using Transformers for RL. In *International Conference on Learning Representations (ICLR)*.

[361] Geunwoo Kim, Pierre Baldi, and Stephen McAleer. 2024. Language models can solve computer tasks. *Advances in Neural Information Processing Systems*, 36.

[362] Kuno Kim, Shivam Garg, Kirankumar Shiragur, and Stefano Ermon. 2021. Reward identification in inverse reinforcement learning. In *International Conference on Machine Learning*, pages 5496–5505. PMLR.

[363] Pieter-Jan Kindermans, Kristof T Schütt, Maximilian Alber, Klaus-Robert Müller, Dumitru Erhan, Been Kim, and Sven Dähne. 2018. Learning how to explain neural networks: Patternnet and patternattribution. In *International Conference on Learning Representations*.

[364] Megan Kinniment, Lucas Jun Koba Sato, Haoxing Du, Brian Goodrich, Max Hasin, Lawrence Chan, Luke Harold Miles, Tao R Lin, Hjalmar Wijk, Joel Burget, Aaron Ho, Elizabeth Barnes, and Paul Christiano. 2023. Evaluating language-model agents on realistic autonomous tasks. https://evals.alignment.org/language-model-pilot-report.

[365] Andrei Kirilenko, Albert S Kyle, Mehrdad Samadi, and Tugkan Tuzun. 2017. The flash crash: High-frequency trading in an electronic market. *The Journal of Finance*, 72(3):967–998.

[366] Svetlana Kiritchenko and Saif M Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508*.

[367] Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. 2017. Self-normalizing neural networks. *Advances in Neural Information Processing Systems*, 30.

[368] Toryn Q Klassen, Sheila A McIlraith, Christian Muise, and Jarvis Xu. 2022. Planning to avoid side effects. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36(9), pages 9830–9839.

[369] Franziska Klügl, Manuel Fehler, and Rainer Herrler. 2005. About the role of the environment in multi-agent simulations. In *Environments for Multi-Agent Systems: First International Workshop, E4MAS 2004, New York, NY, July 19, 2004, Revised Selected Papers 1*, pages 127–149. Springer.

[370] W Bradley Knox, Alessandro Allievi, Holger Banzhaf, Felix Schmitt, and Peter Stone. 2023. Reward (mis) design for autonomous driving. *Artificial Intelligence*, 316:103829.

[371] W Bradley Knox and Peter Stone. 2008. Tamer: Training an agent manually via evaluative reinforcement. In *2008 7th IEEE international conference on development and learning*, pages 292–297. IEEE.

[372] W Bradley Knox and Peter Stone. 2012. Reinforcement learning from simultaneous human and mdp reward. In *AAMAS*, volume 1004, pages 475–482. Valencia.

[373] W Bradley Knox and Peter Stone. 2013. Learning non-myopically from human-generated reward. In *Proceedings of the 2013 international conference on Intelligent user interfaces*, pages 191–202.

[374] W Bradley Knox, Peter Stone, and Cynthia Breazeal. 2013. Training a robot via human feedback: A case study. In *Social Robotics: 5th International Conference, ICSR 2013, Bristol, UK, October 27-29, 2013, Proceedings 5*, pages 460–470. Springer.

[375] William Bradley Knox. 2012. Learning from human-generated reward.

[376] Leonie Koessler and Jonas Schuett. 2023. Risk assessment at agi companies: A review of popular risk assessment techniques from other safety-critical industries. *arXiv preprint arXiv:2307.08823*.

[377] Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR.

[378] Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, et al. 2024. Openassistant conversations-democratizing large language model alignment. *Advances in Neural Information Processing Systems*, 36.

[379] Tomasz Korbak, Kejian Shi, Angelica Chen, Rasika Vinayak Bhalerao, Christopher Buckley, Jason Phang, Samuel R Bowman, and Ethan Perez. 2023. Pretraining language models with human preferences. In *International Conference on Machine Learning*, pages 17506–17533. PMLR.

[380] Peter Krafft, Chris Baker, Alex Pentland, and Joshua Tenenbaum. 2016. Modeling human ad hoc coordination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30–(1).

[381] Victoria Krakovna. 2020. More instances about specification gaming. https://docs.google.com/spreadsheets/d/e/2PACX-1vRPiprOaC3HsCf5Tuum8bRfzYUiKLRqJmbOoC-32JorNdfyTiRRsR7Ea5eWtvsWzuxo8bjOxCG84dAg/pubhtml.

[382] Victoria Krakovna. 2022. Paradigms of ai alignment: components and enablers. https://vkrakovna.wordpress.com/2022/06/02/paradigms-of-ai-alignment-components-and-enablers.

[383] Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2021. Gedi: Generative discriminator guided sequence generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4929–4952.

[384] Satyapriya Krishna, Tessa Han, Alex Gu, Javin Pombra, Shahin Jabbari, Steven Wu, and Himabindu Lakkaraju. 2022. The disagreement problem in explainable machine learning: A practitioner's perspective. *arXiv preprint arXiv:2202.01602*.

[385] Maya Krishnan. 2020. Against interpretability: a critical examination of the interpretability problem in machine learning. *Philosophy & Technology*, 33(3):487–502.

[386] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. 2021. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pages 5815–5826. PMLR.

[387] David Krueger, Tegan Maharaj, and Jan Leike. 2020. Hidden incentives for auto-induced distributional shift. *arXiv preprint arXiv:2009.09153*.

[388] Andras Kupcsik, Marc Deisenroth, Jan Peters, and Gerhard Neumann. 2013. Data-efficient generalization of robot skills with contextual policy search. In *Proceedings of the AAAI conference on artificial intelligence*, volume 27(1), pages 1401–1407.

[389] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2019. An evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1902.00006*.

[390] Isaac Lage, Andrew Ross, Samuel J Gershman, Been Kim, and Finale Doshi-Velez. 2018. Human-in-the-loop interpretability prior. *Advances in Neural Information Processing Systems*, 31.

[391] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. 2017. Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253.

[392] Richard N Landers and Tara S Behrend. 2023. Auditing the ai auditors: A framework for evaluating fairness and bias in high stakes ai predictive models. *American Psychologist*, 78(1):36.

[393] Chan Lawrence, Adria Garriga-alonso, Nicholas Goldowsky-Dill, Ryan Greenblatt, Jenny Nitishinskaya, Ansh Radhakrishnan, Buck Shlegeris, and Thomas Nate. 2023. Causal Scrubbing: a method for rigorously testing interpretability hypotheses [Redwood Research]. https://www.alignmentforum.org/posts/JvZhhzycHu2Yd57RN/causal-scrubbing-a-method-for-rigorously-testing.

[394] Yann LeCun. 2022. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62.

[395] Deokjae Lee, Seungyong Moon, Junhyeok Lee, and Hyun Oh Song. 2022. Query-efficient and scalable black-box adversarial attacks on discrete sequential data via bayesian optimization. In *International Conference on Machine Learning*, pages 12478–12497. PMLR.

[396] Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. 2023a. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*.

[397] Jaesong Lee, Joong-Hwi Shin, and Jun-Seok Kim. 2017. Interactive visualization and manipulation of attention-based neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 121–126.

[398] Lisa Lee, Benjamin Eysenbach, Emilio Parisotto, Eric Xing, Sergey Levine, and Ruslan Salakhutdinov. 2019. Efficient exploration via state marginal matching. *arXiv preprint arXiv:1906.05274*.

[399] Sunok Lee, Minha Lee, and Sangsu Lee. 2023b. What if artificial intelligence become completely ambient in our daily lives? exploring future human-ai interaction through high fidelity illustrations. *International Journal of Human–Computer Interaction*, 39(7):1371–1389.

[400] Joel Lehman, Jeff Clune, Dusan Misevic, Christoph Adami, Lee Altenberg, Julie Beaulieu, Peter J Bentley, Samuel Bernard, Guillaume Beslon, David M Bryson, et al. 2020. The surprising creativity of digital evolution: A collection of anecdotes from the evolutionary computation and artificial life research communities. *Artificial life*, 26(2):274–306.

[401] Joel Lehman, Kenneth O Stanley, et al. 2008. Exploiting open-endedness to solve problems through the search for novelty. In *ALIFE*, pages 329–336.

[402] Joel Z Leibo, Edgar A Dueñez-Guzman, Alexander Vezhnevets, John P Agapiou, Peter Sunehag, Raphael Koster, Jayd Matyas, Charlie Beattie, Igor Mordatch, and Thore Graepel. 2021. Scalable evaluation of multi-agent reinforcement learning with melting pot. In *International conference on machine learning*, pages 6187–6199. PMLR.

[403] Jan Leike. 2022. A proposal for improving societys values.

[404] Jan Leike. 2023a. Combining weak-to-strong generalization with scalable oversight. https://aligned.substack.com/p/combining-w2sg-with-scalable-oversight?utm_source=post-email-title&publication_id=328633&post_id=139945470&utm_campaign=email-post-title&isFreemail=true&r=2xbqf0.

[405] Jan Leike. 2023b. A proposal for importing society's values. https://aligned.substack.com/p/a-proposal-for-importing-societys-values.

[406] Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. 2018. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*.

[407] Filippa Lentzos. 2022. Ai and biological weapons. In *Armament, Arms Control and Artificial Intelligence: The Janus-faced Nature of Machine Learning in the Military Realm*, pages 91–100. Springer.

[408] Belinda Z Li, Maxwell Nye, and Jacob Andreas. 2021a. Implicit representations of meaning in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1813–1827.

[409] Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, Jiquan Pei, Jinfeng Yi, and Bowen Zhou. 2023a. Trustworthy ai: From principles to practices. *ACM Computing Surveys (CSUR)*, 55(9):1–46.

[410] Chao Li, Kelu Yao, Jin Wang, Boyu Diao, Yongjun Xu, and Quanshi Zhang. 2022a. Interpretable generative adversarial networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36(2), pages 1280–1288.

[411] Jiawei Li, Yiming Li, Xingchun Xiang, Shu-Tao Xia, Siyi Dong, and Yun Cai. 2020. Tnt: An interpretable tree-network-tree learning framework using knowledge distillation. *Entropy*, 22(11):1203.

[412] Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2022b. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *The Eleventh International Conference on Learning Representations*.

[413] Mengxi Li, Alper Canberk, Dylan P Losey, and Dorsa Sadigh. 2021b. Learning human objectives from sequences of physical corrections. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2877–2883. IEEE.

[414] Tao Li and Suresh P Sethi. 2017. A review of dynamic stackelberg game models. *Discrete & Continuous Dynamical Systems-B*, 22(1):125.

[415] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305.

[416] Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning*, pages 6565–6576. PMLR.

References

[417] Tom Lieberum, Matthew Rahtz, János Kramár, Neel Nanda, Geoffrey Irving, Rohin Shah, and Vladimir Mikulik. 2023. Does circuit analysis interpretability scale? evidence from multiple choice capabilities in chinchilla.

[418] Wim BG Liebrand. 1984. The effect of social motives, communication and group size on behaviour in an n-person multi-stage mixed-motive game. *European journal of social psychology*, 14(3):239–264.

[419] Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. Kagnet: Knowledge-aware graph networks for commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2829–2839.

[420] Fengming Lin, Xiaolei Fang, and Zheming Gao. 2022a. Distributionally robust optimization: A review on theory and applications. *Numerical Algebra, Control and Optimization*, 12(1):159–212.

[421] Jessy Lin, Daniel Fried, Dan Klein, and Anca Dragan. 2022b. Inferring rewards from language in context. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8546–8560.

[422] Stephanie Lin, Jacob Hilton, and Owain Evans. 2022c. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252.

[423] Yen-Ting Lin and Yun-Nung Chen. 2023. LLM-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, pages 47–58.

[424] Zachary C Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57.

[425] Qin Liu, Meng Zheng, Benjamin Planche, Srikrishna Karanam, Terrence Chen, Marc Niethammer, and Ziyan Wu. 2022. Pseudoclick: Interactive Image Segmentation with Click Imitation. In *European Conference on Computer Vision (ECCV)*.

[426] Ruibo Liu, Ruixin Yang, Chenyan Jia, Ge Zhang, Diyi Yang, and Soroush Vosoughi. 2024a. Training socially aligned language models on simulated social interactions. In *The Twelfth International Conference on Learning Representations*.

[427] Shusen Liu, Tao Li, Zhimin Li, Vivek Srikumar, Valerio Pascucci, and Peer-Timo Bremer. 2018. Visual interrogation of attention-based models for natural language inference and machine comprehension. Technical report, Lawrence Livermore National Lab.(LLNL), Livermore, CA (United States).

[428] Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. 2024b. Agentbench: Evaluating LLMs as agents. In *The Twelfth International Conference on Learning Representations*.

[429] Xiaodong Liu, Hao Cheng, Pengcheng He, Weizhu Chen, Yu Wang, Hoifung Poon, and Jianfeng Gao. 2020. Adversarial training for large neural language models. *arXiv preprint arXiv:2004.08994*.

[430] Yi Liu, Gelei Deng, Zhengzi Xu, Yuekang Li, Yaowen Zheng, Ying Zhang, Lida Zhao, Tianwei Zhang, and Yang Liu. 2023. Jailbreaking chatgpt via prompt engineering: An empirical study. *arXiv preprint arXiv:2305.13860*.

[431] Robert Loftin, Bei Peng, James MacGlashan, Michael L Littman, Matthew E Taylor, Jeff Huang, and David L Roberts. 2016. Learning behaviors via human-delivered discrete feedback: modeling implicit feedback strategies to speed up learning. *Autonomous Agents and Multi-Agent Systems*, 30:30–59.

[432] Dylan P Losey, Andrea Bajcsy, Marcia K O'Malley, and Anca D Dragan. 2022. Physical interaction as communication: Learning robot objectives online from human corrections. *The International Journal of Robotics Research*, 41(1):20–44.

[433] Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in Neural Information Processing Systems*, 30.

[434] Ekdeep Singh Lubana, Eric J Bigelow, Robert P Dick, David Krueger, and Hidenori Tanaka. 2023. Mechanistic mode connectivity. In *International Conference on Machine Learning*, pages 22965–23004. PMLR.

[435] Daniel D Lundstrom, Tianjian Huang, and Meisam Razaviyayn. 2022. A rigorous study of integrated gradients method and extensions to internal neuron attributions. In *International Conference on Machine Learning*, pages 14485–14508. PMLR.

[436] Corey Lynch, Mohi Khansari, Ted Xiao, Vikash Kumar, Jonathan Tompson, Sergey Levine, and Pierre Sermanet. 2020. Learning latent plans from play. In *Conference on robot learning*, pages 1113–1132. PMLR.

[437] Corey Lynch, Ayzaan Wahid, Jonathan Tompson, Tianli Ding, James Betker, Robert Baruch, Travis Armstrong, and Pete Florence. 2023. Interactive language: Talking to robots in real time. *IEEE Robotics and Automation Letters*.

[438] Weiqin Ma, Pu Duan, Sanmin Liu, Guofei Gu, and Jyh-Charn Liu. 2012. Shadow attacks: automatically evading system-call-behavior based malware detection. *Journal in Computer Virology*, 8:1–13.

[439] Zixian Ma, Rose Wang, Fei-Fei Li, Michael Bernstein, and Ranjay Krishna. 2022. Elign: Expectation alignment as a multi-agent intrinsic reward. *Advances in Neural Information Processing Systems*, 35:8304–8317.

[440] Matthijs M Maas. 2021. Aligning ai regulation to sociotechnical change. *Oxford Handbook on AI Governance (Oxford University Press, 2022 forthcoming)*.

[441] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

[442] James MacGlashan, Mark K Ho, Robert Loftin, Bei Peng, Guan Wang, David L Roberts, Matthew E Taylor, and Michael L Littman. 2017. Interactive learning from policy-dependent human feedback. In *International conference on machine learning*, pages 2285–2294. PMLR.

[443] Alasdair MacIntyre. 2013. *After virtue*. A&C Black.

[444] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*.

[445] Andreas Madsen, Siva Reddy, and Sarath Chandar. 2022. Post-hoc interpretability for neural nlp: A survey. *ACM Computing Surveys (CSUR)*, 55(8):1–42.

[446] MAI. 2023. Introducing democratic fine-tuning.

[447] Daniel J Mankowitz, Andrea Michi, Anton Zhernov, Marco Gelmi, Marco Selvi, Cosmin Paduraru, Edouard Leurent, Shariq Iqbal, Jean-Baptiste Lespiau, Alex Ahern, et al. 2023. Faster sorting algorithms discovered using deep reinforcement learning. *Nature*, 618(7964):257–263.

[448] Aaron Mannes. 2020. Governance, risk, and artificial intelligence. *AI Magazine*, 41(1):61–69.

[449] James Manyika, Michael Chui, Mehdi Miremadi, Jacques Bughin, Katy George, Paul Willmott, and Martin Dewhurst. 2017. A future that works: Ai, automation, employment, and productivity. *McKinsey Global Institute Research, Tech. Rep*, 60:1–135.

[450] Xiaofeng Mao, Yuefeng Chen, Ranjie Duan, Yao Zhu, Gege Qi, Xiaodan Li, Rong Zhang, Hui Xue, et al. 2022. Enhance the visual representation via discrete adversarial training. *Advances in Neural Information Processing Systems*, 35:7520–7533.

[451] Peter Marbach and John N Tsitsiklis. 2001. Simulation-based optimization of markov reward processes. *IEEE Transactions on Automatic Control*, 46(2):191–209.

[452] Gary Marcus. 2018. Deep learning: A critical appraisal.

[453] David Mascharka, Philip Tran, Ryan Soklaski, and Arjun Majumdar. 2018. Transparency by design: Closing the gap between performance and interpretability in visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4942–4950.

[454] Charles G McClintock and Eddy Van Avermaet. 1982. Social values and rules of fairness: A theoretical perspective. *Cooperation and helping behavior*, pages 43–71.

[455] Thomas McGrath, Matthew Rahtz, Janos Kramar, Vladimir Mikulik, and Shane Legg. 2023. The hydra effect: Emergent self-repair in language model computations. *arXiv preprint arXiv:2307.15771*.

[456] Kevin R McKee, Ian Gemp, Brian McWilliams, Edgar A Duèñez-Guzmán, Edward Hughes, and Joel Z Leibo. 2020. Social diversity and social preferences in mixed-motive reinforcement learning. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, pages 869–877.

[457] Lev E McKinney, Yawen Duan, David Krueger, and Adam Gleave. 2022. On the fragility of learned reward functions. In *NeurIPS ML Safety Workshop*.

[458] Scott McLean, Gemma JM Read, Jason Thompson, Chris Baber, Neville A Stanton, and Paul M Salmon. 2023. The risks associated with artificial general intelligence: A systematic review. *Journal of Experimental & Theoretical Artificial Intelligence*, 35(5):649–663.

[459] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35.

[460] Bahar Memarian and Tenzin Doleck. 2023. Fairness, accountability, transparency, and ethics (fate) in artificial intelligence (ai), and higher education: A systematic review. *Computers and Education: Artificial Intelligence*, page 100152.

[461] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022a. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.

[462] Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. 2022b. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations*.

[463] Alex Mennen. 2018. A comment on the ida-alphagozero metaphor; capabilities versus alignment. `https://www.alignmentforum.org/posts/yXFKh2jGysQNfX2NM/a-comment-on-the-ida-alphagozero-metaphor-capabilities`.

[464] Bruno Mermet and Gaële Simon. 2016. Formal verication of ethical properties in multiagent systems. In *1st Workshop on Ethics in the Design of Intelligent Agents*.

[465] David M Messick and Charles G McClintock. 1968. Motivational bases of choice in experimental games. *Journal of experimental social psychology*, 4(1):1–25.

[466] Meta. 2023. Meta and microsoft introduce the next generation of llama. `https://ai.meta.com/blog/llama-2`.

[467] Julian Michael, Ari Holtzman, Alicia Parrish, Aaron Mueller, Alex Wang, Angelica Chen, Divyam Madaan, Nikita Nangia, Richard Yuanzhe Pang, Jason Phang, et al. 2022. What do nlp researchers believe? results of the nlp community metasurvey. *arXiv preprint arXiv:2208.12852*.

[468] Michaelcohen. 2020. the-ai-debate-debate. `https://www.alignmentforum.org/posts/L3QDs6of4Rb2TgpRD/the-ai-debate-debate`.

[469] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

[470] Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys (CSUR)*, 56(2):1–40.

[471] Yisroel Mirsky, Ambra Demontis, Jaidip Kotak, Ram Shankar, Deng Gelei, Liu Yang, Xiangyu Zhang, Maura Pintor, Wenke Lee, Yuval Elovici, et al. 2023. The threat of offensive ai to organizations. *Computers & Security*, 124:103006.

[472] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.

[473] Thomas M Moerland, Joost Broekens, Aske Plaat, Catholijn M Jonker, et al. 2023. Model-based reinforcement learning: A survey. *Foundations and Trends® in Machine Learning*, 16(1):1–118.

[474] Sina Mohseni, Niloofar Zarei, and Eric D Ragan. 2021. A multidisciplinary survey and framework for design and evaluation of explainable ai systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 11(3-4):1–45.

[475] Jakob Mökander, Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi. 2023. Auditing large language models: a three-layered approach. *AI and Ethics*, pages 1–31.

[476] Ari S Morcos, David GT Barrett, Neil C Rabinowitz, and Matthew Botvinick. 2018. On the importance of single directions for generalization. In *International Conference on Learning Representations*.

[477] Alexander Mordvintsev, Chris Olah, and Mike Tyka. 2015. Inceptionism: Going deeper into neural networks. *Google Research Blog*.

References

[478] Emad Mostaque. 2022. Democratizing ai, stable diffusion & generative models. https://exchange.scale.com/public/videos/emad-mostaque-stability-ai-stable-diffusion-open-source.

[479] Darius Muglich, Christian Schroeder de Witt, Elise van der Pol, Shimon Whiteson, and Jakob Foerster. 2022. Equivariant networks for zero-shot coordination. *Advances in Neural Information Processing Systems*, 35:6410–6423.

[480] Gabriel Mukobi. 2022. Iterated distillation-amplification, gato, and proto-agi. https://www.lesswrong.com/posts/Evyk8eb6b7tFd6pxJ/iterated-distillation-amplification-gato-and-proto-agi-re.

[481] Arslan Munir, Alexander Aved, and Erik Blasch. 2022. Situational awareness: techniques, challenges, and prospects. *AI*, 3(1):55–77.

[482] Kevin P. Murphy. 2023. *Probabilistic Machine Learning: Advanced Topics*. MIT Press.

[483] Ryan O Murphy and Kurt A Ackermann. 2014. Social value orientation: Theoretical and measurement issues in the study of social preferences. *Personality and Social Psychology Review*, 18(1):13–41.

[484] Ryan O Murphy, Kurt A Ackermann, and Michel JJ Handgraaf. 2011. Measuring social value orientation. *Judgment and Decision making*, 6(8):771–781.

[485] Grazia Murtarelli, Anne Gregory, and Stefania Romenti. 2021. A conversation-based perspective for shaping ethical human–machine interactions: The particular challenge of chatbots. *Journal of Business Research*, 129:927–935.

[486] Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. Stereoset: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371.

[487] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

[488] Neel Nanda. 2023a. Attribution patching: Activation patching at industrial scale.

[489] Neel Nanda. 2023b. Othello-gpt: Future work i am excited about. https://www.alignmentforum.org/posts/qgK7smTvJ4DB8rZ6h/othello-gpt-future-work-i-am-excited-about.

[490] Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2022. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations*.

[491] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel Bowman. 2020. Crows-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967.

[492] Ashutosh Nayyar, Aditya Mahajan, and Demosthenis Teneketzis. 2013. Decentralized stochastic control with partial history sharing: A common information approach. *IEEE Transactions on Automatic Control*, 58(7):1644–1658.

[493] Michael Neely, Stefan F Schouten, Maurits JR Bleeker, and Ana Lucic. 2021. Order in the court: Explainable ai methods prone to disagreement. *arXiv preprint arXiv:2105.03287*.

[494] Andrew Y Ng, Daishi Harada, and Stuart Russell. 1999. Policy invariance under reward transformations: Theory and application to reward shaping. In *Icml*, volume 99, pages 278–287. Citeseer.

[495] Andrew Y Ng, Stuart Russell, et al. 2000. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, page 2.

[496] Richard Ngo. 2020a. Agi safety from first principles. https://www.alignmentforum.org/s/mzgtmmTKKn5MuCzFJ.

[497] Richard Ngo. 2020b. continuing-the-takeoffs-debate. https://www.alignmentforum.org/posts/Tpn2Fx9daLvj28kes/continuing-the-takeoffs-debate.

[498] Richard Ngo. 2021. /why i m excited about debate. https://www.alignmentforum.org/posts/LDsSqXf9Dpu3J3gHD/why-i-m-excited-about-debate.

[499] Richard Ngo, Lawrence Chan, and Sören Mindermann. 2024. The alignment problem from a deep learning perspective: A position paper. In *The Twelfth International Conference on Learning Representations*.

[500] Anh Nguyen, Jason Yosinski, and Jeff Clune. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[501] Anh Nguyen, Jason Yosinski, and Jeff Clune. 2016. Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks. *arXiv preprint arXiv:1602.03616*.

[502] Chi Nguyen. 2020. My understanding of paul christiano's iterated amplification al safety research agenda. https://www.alignmentforum.org/posts/PT8vSxsusqWuN7JXp/my-understanding-of-paul-christiano-s-iterated-amplification#A_mathematical_way_of_solving_Go_is_impossible.

[503] Tianwei Ni, Harshit Sikchi, Yufei Wang, Tejus Gupta, Lisa Lee, and Ben Eysenbach. 2021. f-irl: Inverse reinforcement learning via state marginal matching. In *Conference on Robot Learning*, pages 529–551. PMLR.

[504] Nassim Nicholas. 2008. The black swan: the impact of the highly improbable. *Journal of the Management Training Institut*, 36(3):56.

[505] Safiya Umoja Noble. 2018. Algorithms of oppression. In *Algorithms of oppression*. New York university press.

[506] Safiya Umoja Noble, Beatrice Dias, Sara Cole Stratton, Aimee van Wynsberghe, Carlos Affonso Souza, Ilene Carpenter, Alvaro Martin Enriquez, and Emily Ratté. 2021. Ai regulation through an intergenerational lens. https://www3.weforum.org/docs/WEF_AI_Regulation_through_an_Intergenerational_Lens_2021.pdf.

[507] Ritesh Noothigattu, Snehalkumar Gaikwad, Edmond Awad, Sohan Dsouza, Iyad Rahwan, Pradeep Ravikumar, and Ariel Procaccia. 2018. A voting-based system for ethical decision making. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

[508] Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, et al. 2020. Bias in data-driven artificial intelligence systems—an introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3):e1356.

[509] OECD. 2019. Oecd principles on artificial intelligence. https://oecd.ai/en/ai-principles.

[510] OecdAI. 2021. Ai principles. https://oecd.ai/en/dashboards/ai-principles/P8.

[511] Caspar Oesterheld. 2021. Approval-directed agency and the decision theory of newcomb-like problems. *Synthese*, 198(Suppl 27):6491–6504.

[512] Caspar Oesterheld and Vincent Conitzer. 2022. Safe pareto improvements for delegated game playing. *Autonomous Agents and Multi-Agent Systems*, 36(2):46.

[513] Chris Olah. 2014. Visualizing mnist: An exploration of dimensionality reduction. https://colah.github.io/posts/2014-10-Visualizing-MNIST.

[514] Chris Olah. 2015. Visualizing representations: Deep learning and human beings. https://colah.github.io/posts/2015-01-Visualizing-Representations.

[515] Chris Olah. 2023. Interpretability dreams. https://transformer-circuits.pub/2023/interpretability-dreams/index.html#larger-scale.

[516] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001.

[517] Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. 2018. The building blocks of interpretability. *Distill*, 3(3):e10.

[518] Chris Olah et al. 2017. Feature visualization. *Distill*.

[519] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. 2022. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*.

[520] Stephen M Omohundro. 2008. The basic ai drives. In *AGI*, volume 171, pages 483–492.

References

[521] OpenAI. 2021a. Curve detectors. `https://distill.pub/2020/circuits/curve-detectors`.

[522] OpenAI. 2021b. Weight banding. `https://distill.pub/2020/circuits/weight-banding`.

[523] OpenAI. 2023a. Gpt-4 technical report.

[524] OpenAI. 2023b. Gpt-4v(ision) system card. `https://cdn.openai.com/papers/GPTV_System_Card.pdf`.

[525] OpenAI. 2023c. Introducing superalignment. `https://openai.com/blog/introducing-superalignment`. Accessed on July 5, 2023.

[526] Robert Opp. 2023. Committing to bridging the digital divide in least developed countries. `https://www.undp.org/blog/committing-bridging-digital-divide-least-developed-countries`.

[527] Afshin Oroojlooy and Davood Hajinezhad. 2023. A review of cooperative multi-agent deep reinforcement learning. *Applied Intelligence*, 53(11):13677–13722.

[528] Takayuki Osa, Joni Pajarinen, Gerhard Neumann, J Andrew Bagnell, Pieter Abbeel, Jan Peters, et al. 2018. An algorithmic perspective on imitation learning. *Foundations and Trends® in Robotics*, 7(1-2):1–179.

[529] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

[530] Lorenzo Pacchiardi, Alex James Chan, Sören Mindermann, Ilan Moscovitz, Alexa Yue Pan, Yarin Gal, Owain Evans, and Jan M. Brauner. 2024. How to catch an AI liar: Lie detection in black-box LLMs by asking unrelated questions. In *The Twelfth International Conference on Learning Representations*.

[531] Malayandi Palan, Gleb Shevchuk, Nicholas Charles Landolfi, and Dorsa Sadigh. 2019. Learning reward functions by integrating human demonstrations and preferences. In *Robotics: Science and Systems*.

[532] Alexander Pan, Kush Bhatia, and Jacob Steinhardt. 2021. The effects of reward misspecification: Mapping and mitigating misaligned models. In *International Conference on Learning Representations*.

[533] Alexander Pan, Jun Shern Chan, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside, Jonathan Ng, Hanlin Zhang, Scott Emmons, and Dan Hendrycks. 2023a. Do the rewards justify the means? measuring trade-offs between rewards and ethical behavior in the machiavelli benchmark. *ICML*.

[534] Alexander Pan, Jun Shern Chan, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside, Hanlin Zhang, Scott Emmons, and Dan Hendrycks. 2023b. Do the rewards justify the means? measuring trade-offs between rewards and ethical behavior in the machiavelli benchmark. In *International Conference on Machine Learning*, pages 26837–26867. PMLR.

[535] Rahul Pandey, Hemant Purohit, Carlos Castillo, and Valerie L Shalin. 2022. Modeling and mitigating human annotation errors to design efficient stream processing systems with human-in-the-loop machine learning. *International Journal of Human-Computer Studies*, 160:102772.

[536] Paulina Karolina Pankowska. 2020. Framework on ethical aspects of artificial intelligence, robotics and related technologies. *European Parliament*.

[537] Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael Wellman. 2016. Towards the science of security and privacy in machine learning. *arXiv preprint arXiv:1611.03814*.

[538] Simone Parisi, Aravind Rajeswaran, Senthil Purushwalkam, and Abhinav Gupta. 2022. The unsurprising effectiveness of pre-trained vision models for control. In *International Conference on Machine Learning*, pages 17359–17371. PMLR.

[539] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023a. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22.

[540] Peter S Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. 2023b. Ai deception: A survey of examples, risks, and potential solutions. *arXiv preprint arXiv:2308.14752*.

[541] Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. Bbq: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105.

References

[542] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11).

[543] Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *International Conference on Learning Representations*.

[544] Hammond Pearce, Baleegh Ahmad, Benjamin Tan, Brendan Dolan-Gavitt, and Ramesh Karri. 2022. Asleep at the keyboard? assessing the security of github copilot's code contributions. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 754–768. IEEE.

[545] Will Pearce and Joseph Lucas. 2023. Nvidia ai red team: An introduction. https://developer.nvidia.com/blog/nvidia-ai-red-team-an-introduction.

[546] Andi Peng, Besmira Nushi, Emre Kiciman, Kori Inkpen, and Ece Kamar. 2022. Investigations of performance and bias in human-ai teamwork in hiring. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36-11, pages 12089–12097.

[547] Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. 2020. Performative prediction. In *International Conference on Machine Learning*, pages 7599–7609. PMLR.

[548] Luís Moniz Pereira, Ari Saptawijaya, Luís Moniz Pereira, and Ari Saptawijaya. 2016a. Bridging two realms of machine ethics. *Programming machine ethics*, pages 159–165.

[549] Luís Moniz Pereira, Ari Saptawijaya, et al. 2016b. *Programming machine ethics*, volume 26. Springer.

[550] Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3419–3448.

[551] Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. 2023. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13387–13434. Association for Computational Linguistics.

[552] Julien Perolat, Joel Z Leibo, Vinicius Zambaldi, Charles Beattie, Karl Tuyls, and Thore Graepel. 2017. A multi-agent reinforcement learning model of common-pool resource appropriation. *Advances in Neural Information Processing Systems*, 30.

[553] Lucas Perry. 2020. Evan hubinger on inner alignment, outer alignment, and proposals for building safe advanced ai. https://www.alignmentforum.org/posts/qZGoHkRgANQpGHWnu/evan-hubinger-on-inner-alignment-outer-alignment-and.

[554] J Peters, Peter Buhlmann, and N Meinshausen. 2015. Causal inference using invariant prediction: identification and confidence intervals. arxiv. *Methodology*.

[555] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2017. *Elements of causal inference: foundations and learning algorithms*. The MIT Press.

[556] Steve Phelps and Yvan I. Russell. 2023. Investigating emergent goal-like behaviour in large language models using experimental economics.

[557] Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. 2017. Robust adversarial reinforcement learning. In *International Conference on Machine Learning*, pages 2817–2826. PMLR.

[558] James Pita, Manish Jain, Milind Tambe, Fernando Ordóñez, and Sarit Kraus. 2010. Robust solutions to stackelberg games: Addressing bounded rationality and limited observations in human cognition. *Artificial Intelligence*, 174(15):1142–1171.

[559] Fabrizio Pittorino, Antonio Ferraro, Gabriele Perugini, Christoph Feinauer, Carlo Baldassi, and Riccardo Zecchina. 2022. Deep networks on toroids: removing symmetries reveals the structure of flat regions in the landscape geometry. In *International Conference on Machine Learning*, pages 17759–17781. PMLR.

[560] Robin L Plackett. 1975. The analysis of permutations. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 24(2):193–202.

[561] Dean A Pomerleau. 1991. Efficient training of artificial neural networks for autonomous navigation. *Neural computation*, 3(1):88–97.

[562] Karl Popper. 2005. *The logic of scientific discovery*. Routledge.

References

[563] Omid Poursaeed, Tianxing Jiang, Harry Yang, Serge Belongie, and Ser-Nam Lim. 2021. Robustness and generalization via generative adversarial training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15711–15720.

[564] Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. 2022. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*.

[565] Lutz Prechelt. 2002. Early stopping-but when? In *Neural Networks: Tricks of the trade*, pages 55–69. Springer.

[566] Dale Purves, George J Augustine, David Fitzpatrick, Lawrence C Katz, Anthony-Samuel LaMantia, James O McNamara, and S Mark. Williams. 2001. *Neuroscience, 2nd edition*. Sinauer Associates.

[567] Martin L Puterman. 2014. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

[568] Tianyi Qiu, Fanzhi Zeng, Jiaming Ji, Dong Yan, Kaile Wang, Jiayi Zhou, Han Yang, Josef Dai, Xuehai Pan, and Yaodong Yang. 2024. Rethinking information structures in rlhf: Reward generalization from a graph theory perspective. *arXiv preprint arXiv:2402.10184*.

[569] R Quian Quiroga, Leila Reddy, Gabriel Kreiman, Christof Koch, and Itzhak Fried. 2005. Invariant visual representation by single neurons in the human brain. *Nature*, 435(7045):1102–1107.

[570] Ansh Radhakrishnan, Karina Nguyen, Anna Chen, Carol Chen, Carson Denison, Danny Hernandez, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, et al. 2023. Question decomposition improves the faithfulness of model-generated reasoning. *arXiv preprint arXiv:2307.11768*.

[571] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

[572] Hamed Rahimian and Sanjay Mehrotra. 2019. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*.

[573] Dheeraj Rajagopal, Vidhisha Balachandran, Eduard H Hovy, and Yulia Tsvetkov. 2021. Selfexplain: A self-explaining architecture for neural text classifiers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 836–850.

[574] Inioluwa Deborah Raji, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily Denton. 2020. Saving face: Investigating the ethical concerns of facial recognition auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 145–151.

[575] Ram Shankar Siva Kumar. 2023. Microsoft ai red team building future of safer ai. https://www.microsoft.com/en-us/security/blog/2023/08/07/microsoft-ai-red-team-building-future-of-safer-ai.

[576] Deepak Ramachandran and Eyal Amir. 2007. Bayesian inverse reinforcement learning. In *IJCAI*, volume 7, pages 2586–2591.

[577] Tilman Räuker, Anson Ho, Stephen Casper, and Dylan Hadfield-Menell. 2023. Toward transparent ai: A survey on interpreting the inner structures of deep neural networks. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 464–483. IEEE.

[578] Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan D Cotterell. 2022. Linear adversarial concept erasure. In *International Conference on Machine Learning*, pages 18400–18421. PMLR.

[579] Harish Ravichandar, Athanasios S Polydoros, Sonia Chernova, and Aude Billard. 2020. Recent advances in robot learning from demonstration. *Annual review of control, robotics, and autonomous systems*, 3:297–330.

[580] Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. 2021. Probing the probing paradigm: Does probing accuracy entail task relevance?

[581] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR.

[582] Siddharth Reddy, Anca Dragan, Sergey Levine, Shane Legg, and Jan Leike. 2020. Learning human objectives by evaluating hypothetical behavior. In *International Conference on Machine Learning*, pages 8020–8029. PMLR.

References

[583] Siddharth Reddy, Anca D Dragan, and Sergey Levine. 2019. Sqil: Imitation learning via reinforcement learning with sparse rewards. In *International Conference on Learning Representations*.

[584] Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gómez Colmenarejo, Alexander Novikov, Gabriel Barth-maron, Mai Giménez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. 2022. A generalist agent. *Transactions on Machine Learning Research*.

[585] Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim. 2019. Visualizing and measuring the geometry of bert. *Advances in Neural Information Processing Systems*, 32.

[586] Yankun Ren, Jianbin Lin, Siliang Tang, Jun Zhou, Shuang Yang, Yuan Qi, and Xiang Ren. 2020. Generating natural language adversarial examples on a large scale with generative models. In *ECAI 2020*, pages 2156–2163. IOS Press.

[587] Richard Ngo. 2022. Gradient hacking. https://www.alignmentforum.org/posts/EeAgytDZb DjRznPMA/gradient-hacking-definitions-and-examples.

[588] Mattia Rigotti, Christoph Miksovic, Ioana Giurgiu, Thomas Gschwind, and Paolo Scotton. 2022. Attention-based interpretability with concept transformers. In *International Conference on Learning Representations*.

[589] Mark Ring and Laurent Orseau. 2011. Delusion, survival, and intelligent agents. In *Artificial General Intelligence: 4th International Conference, AGI 2011, Mountain View, CA, USA, August 3-6, 2011. Proceedings 4*, pages 11–20. Springer.

[590] Yuji Roh, Geon Heo, and Steven Euijong Whang. 2019. A survey on data collection for machine learning: a big data-ai integration perspective. *IEEE Transactions on Knowledge and Data Engineering*, 33(4):1328–1347.

[591] Milton Rokeach. 1973. *The nature of human values*. Free press.

[592] Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Marcos Zampieri, and Preslav Nakov. 2021. Solid: A large-scale semi-supervised dataset for offensive language identification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 915–928.

[593] Andrew Ross, Isaac Lage, and Finale Doshi-Velez. 2017. The neural lasso: Local linear sparsity for interpretable explanations. In *Workshop on Transparent and Interpretable Machine Learning in Safety Critical Environments, 31st Conference on Neural Information Processing Systems*, volume 4.

[594] Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. 2011. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 627–635. JMLR Workshop and Conference Proceedings.

[595] Francesca Rossi, Kristen Brent Venable, and Toby Walsh. 2011. *A Short Introduction to Preferences: Between AI and Social Choice*. Morgan & Claypool Publishers.

[596] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. 2004. " grabcut" interactive foreground extraction using iterated graph cuts. *ACM transactions on graphics (TOG)*, 23(3):309–314.

[597] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

[598] Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14.

[599] Tim Rudner and Helen Toner. 2021a. Key concepts in ai safety: Interpretability in machine learning. http s://cset.georgetown.edu/publication/key-concepts-in-ai-safety-interpretab ility-in-machine-learning.

[600] Tim Rudner and Helen Toner. 2021b. Key concepts in ai safety: Robustness and adversarial examples. https://cset.georgetown.edu/publication/key-concepts-in-ai-safety-robustn ess-and-adversarial-examples.

[601] Stuart Russell. 2019. *Human compatible: Artificial intelligence and the problem of control*. Penguin.

[602] Stuart Russell, Daniel Dewey, and Max Tegmark. 2015. Research priorities for robust and beneficial artificial intelligence. *AI Magazine*, 36(4):105–114.

[603] Noveen Sachdeva and Julian McAuley. 2023. Data distillation: A survey. *Transactions on Machine Learning Research*. Survey Certification.

[604] Joel L Sachs, Ulrich G Mueller, Thomas P Wilcox, and James J Bull. 2004. The evolution of cooperation. *The Quarterly review of biology*, 79(2):135–160.

[605] Dorsa Sadigh, Anca D Dragan, Shankar Sastry, and Sanjit A Seshia. 2017. *Active preference-based learning of reward functions*. Escholarship.

[606] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. 2020. Distributionally robust neural networks. In *International Conference on Learning Representations*.

[607] Ananya B Sai, Akash Kumar Mohankumar, and Mitesh M Khapra. 2022. A survey of evaluation metrics used for nlg systems. *ACM Computing Surveys (CSUR)*, 55(2):1–39.

[608] Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T Rodolfa, and Rayid Ghani. 2018. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577*.

[609] Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller. 2019. *Explainable AI: interpreting, explaining and visualizing deep learning*, volume 11700. Springer Nature.

[610] Jonas B Sandbrink. 2023. Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools. *arXiv preprint arXiv:2306.13952*.

[611] Lindsay Sanneman and Julie A Shah. 2020. A situation awareness-based framework for design and evaluation of explainable ai. In *Explainable, Transparent Autonomous Agents and Multi-Agent Systems: Second International Workshop, EXTRAAMAS 2020, Auckland, New Zealand, May 9–13, 2020, Revised Selected Papers 2*, pages 94–110. Springer.

[612] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR.

[613] Beatrice Dias Sara Stratton. 2021. Why we must consider the intergenerational impacts of ai. https://www.weforum.org/agenda/2021/10/why-we-must-consider-the-intergenerational-impact-of-ai.

[614] Fumihiro Sasaki and Ryota Yamashina. 2020. Behavioral cloning from noisy demonstrations. In *International Conference on Learning Representations*.

[615] William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. 2022. Self-critiquing models for assisting human evaluators. *arXiv preprint arXiv:2206.05802*.

[616] Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C Parkes, and Yang Liu. 2019. How do fairness definitions fare? examining public attitudes towards algorithmic definitions of fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 99–106.

[617] Stefan Schaal. 1996. Learning from demonstration. *Advances in Neural Information Processing Systems*, 9.

[618] Stefan Schaal. 1999. Is imitation learning the route to humanoid robots? *Trends in cognitive sciences*, 3(6):233–242.

[619] Nino Scherrer, Claudia Shi, Amir Feder, and David Blei. 2024. Evaluating the moral beliefs encoded in llms. *Advances in Neural Information Processing Systems*, 36.

[620] Johannes Schneider and Michalis Vlachos. 2021. Explaining neural networks by decoding layer activations. In *Advances in Intelligent Data Analysis XIX: 19th International Symposium on Intelligent Data Analysis, IDA 2021, Porto, Portugal, April 26–28, 2021, Proceedings 19*, pages 63–75. Springer.

[621] Jonas Schuett, Noemi Dreksler, Markus Anderljung, David McCaffary, Lennart Heim, Emma Bluemke, and Ben Garfinkel. 2023. Towards best practices in agi safety and governance: A survey of expert opinion. *arXiv preprint arXiv:2305.07153*.

[622] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

[623] Peter Schuster and Karl Sigmund. 1983. Replicator dynamics. *Journal of theoretical biology*, 100(3):533–538.

[624] Shalom H Schwartz. 1992. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In *Advances in experimental social psychology*, volume 25, pages 1–65. Elsevier.

References

[625] Shalom H Schwartz. 1994. Are there universal aspects in the structure and contents of human values? *Journal of social issues*, 50(4):19–45.

[626] Elizabeth Seger, Noemi Dreksler, Richard Moulange, Emily Dardaman, Jonas Schuett, K Wei, Christoph Winter, Mackenzie Arnold, Seán Ó hÉigeartaigh, Anton Korinek, et al. 2023. Open-sourcing highly capable foundation models: An evaluation of risks, benefits, and alternative methods for pursuing open-source objectives. *arXiv preprint arXiv:2311.09227*.

[627] Charbel-Raphael Segerie. 2023. Against almost every theory of impact of interpretability. `https://www.alignmentforum.org/posts/LNA8mubrByG7SFacm/against-almost-every-theory-of-impact-of-interpretability-1`.

[628] Amartya Sen. 1986. Social choice theory. *Handbook of mathematical economics*, 3:1073–1181.

[629] Egemen Sert, Yaneer Bar-Yam, and Alfredo J Morales. 2020. Segregation dynamics with reinforcement learning and agent based modeling. *Scientific Reports*, 10(1):11771.

[630] Rohin Shah, Pedro Freire, Neel Alex, Rachel Freedman, Dmitrii Krasheninnikov, Lawrence Chan, Michael D Dennis, Pieter Abbeel, Anca Dragan, and Stuart Russell. 2020. Benefits of assistance over reward learning. *https://openreview.net/forum?id=DFIoGDZejIB*.

[631] Rohin Shah and Vikrant Varma. 2022. More examples of gmg. `https://www.alignmentforum.org/posts/Cfe2LMmQC4hHTDZ8r/more-examples-of-goal-misgeneralization`.

[632] Rusheb Shah, Quentin Feuillade Montixi, Soroush Pour, Arush Tagade, and Javier Rando. 2023. Scalable and transferable black-box jailbreaks for language models via persona modulation. In *Socially Responsible Language Modelling Research*.

[633] Ali Shahin Shamsabadi, Ricardo Sanchez-Matilla, and Andrea Cavallaro. 2020. Colorfool: Semantic adversarial colorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1151–1160.

[634] Jacob N Shapiro and David A Siegel. 2010. Is this paper dangerous? balancing secrecy and openness in counterterrorism. *Security Studies*, 19(1):66–98.

[635] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Esin DURMUS, Zac Hatfield-Dodds, Scott R Johnston, Shauna M Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2024. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations*.

[636] Pratyusha Sharma, Balakumar Sundaralingam, Valts Blukis, Chris Paxton, Tucker Hermans, Antonio Torralba, Jacob Andreas, and Dieter Fox. 2022. Correcting robot plans with natural language feedback. *arXiv preprint arXiv:2204.05186*.

[637] Kenneth Shaw, Shikhar Bahl, and Deepak Pathak. 2023. Videodex: Learning dexterity from internet videos. In *Conference on Robot Learning*, pages 654–665. PMLR.

[638] Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. 2023. "do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. *arXiv preprint arXiv:2308.03825*.

[639] Amit Sheth, Valerie L Shalin, and Ugur Kursuncu. 2022. Defining and detecting toxicity on social media: context and knowledge are key. *Neurocomputing*, 490:312–318.

[640] Toby Shevlane, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, Nahema Marchal, Markus Anderljung, Noam Kolt, et al. 2023. Model evaluation for extreme risks. *arXiv preprint arXiv:2305.15324*.

[641] Adam Shimi. 2022. How to diversify conceptual alignment: the model behind refine. `https://www.alignmentforum.org/posts/5uiQkyKdejX3aEHLM/how-to-diversify-conceptual-alignment-the-model-behind`.

[642] Daniel Shin, Anca D. Dragan, and Daniel S. Brown. 2023. Benchmarks and algorithms for offline preference-based reward learning. *Trans. Mach. Learn. Res.*, 2023.

[643] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235.

[644] Buck Shlegeris and Ryan Greenblatt. 2023. Meta-level adversarial evaluation of oversight techniques might allow robust measurement of their adequacy. In *Alignment Forum*.

[645] Wai Man Si, Michael Backes, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, Savvas Zannettou, and Yang Zhang. 2022. Why so toxic? measuring and triggering toxic behavior in open-domain chatbots. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 2659–2673.

[646] Harshit Sikchi, Qinqing Zheng, Amy Zhang, and Scott Niekum. 2023. Dual rl: Unification and new methods for reinforcement and imitation learning. In *Sixteenth European Workshop on Reinforcement Learning*.

[647] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489.

[648] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. 2017. Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359.

[649] David Silver, Satinder Singh, Doina Precup, and Richard S Sutton. 2021. Reward is enough. *Artificial Intelligence*, 299:103535.

[650] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

[651] Amanpreet Singh, Tushar Jain, and Sainbayar Sukhbaatar. 2018. Learning when to communicate at scale in multiagent cooperative and competitive tasks. In *International Conference on Learning Representations*.

[652] Munindar P Singh. 2014. Norms as a basis for governing sociotechnical systems. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(1):1–23.

[653] Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. 2022. Defining and characterizing reward gaming. *Advances in Neural Information Processing Systems*, 35:9460–9471.

[654] Joar Max Viktor Skalse, Matthew Farrugia-Roberts, Stuart Russell, Alessandro Abate, and Adam Gleave. 2023. Invariance in policy optimisation and partial identifiability in reward learning. In *International Conference on Machine Learning*, pages 32033–32058. PMLR.

[655] Nate Soares. 2018. The value learning problem. In *Artificial intelligence safety and security*, pages 89–97. Chapman and Hall/CRC.

[656] Nate Soares and Benja Fallenstein. 2014. Aligning superintelligence with human interests: A technical research agenda. *Machine Intelligence Research Institute (MIRI) technical report*, 8.

[657] Nate Soares, Benja Fallenstein, Stuart Armstrong, and Eliezer Yudkowsky. 2015. Corrigibility. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*.

[658] Nate Soares and Benya Fallenstein. 2017. Agent foundations for aligning machine intelligence with human interests: a technical research agenda. *The technological singularity: Managing the journey*, pages 103–125.

[659] Emily H. Soice, Rafael Rocha, Kimberlee Cordova, Michael Specter, and Kevin M. Esvelt. 2023. Can large language models democratize access to dual-use biotechnology?

[660] Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

[661] Jiaming Song, Hongyu Ren, Dorsa Sadigh, and Stefano Ermon. 2018a. Multi-agent generative adversarial imitation learning. *Advances in Neural Information Processing Systems*, 31.

[662] Yang Song, Rui Shu, Nate Kushman, and Stefano Ermon. 2018b. Constructing unrestricted adversarial examples with generative models. *Advances in Neural Information Processing Systems*, 31.

[663] Giovanni Spitale, Nikola Biller-Andorno, and Federico Germani. 2023. Ai model gpt-3 (dis)informs us better than humans. *Science Advances*, 9(26):eadh1850.

[664] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

[665] Bernd Carsten Stahl and Tonii Leach. 2023. Assessing the ethical and social concerns of artificial intelligence in neuroinformatics research: An empirical test of the european union assessment list for trustworthy ai (altai). *AI and Ethics*, 3(3):745–767.

[666] Zach Stein-Perlman, Benjamin Weinstein-Raun, and Katja Grace. 2022. expert survey on progress in ai. *AI Impacts. Available online at: https://aiimpacts. org/2022-expert-survey-on-progress-in-ai (accessed December 7, 2022).*

[667] Jacob Steinhardt. 2015. https://jsteinhardt.wordpress.com/2015/06/24/long-term-and-short-term-challenges-to-ensuring-the-safety-of-ai-systems, title = Long-Term and Short-Term Challenges to Ensuring the Safety of AI Systems.

[668] Jacob Steinhardt and Helen Toner. 2020. Why robustness is key to deploying ai. https://www.brookings.edu/articles/why-robustness-is-key-to-deploying-ai.

[669] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

[670] Peter Stone, Gal Kaminka, Sarit Kraus, and Jeffrey Rosenschein. 2010. Ad hoc autonomous agent teams: Collaboration without pre-coordination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 24, pages 1504–1509.

[671] Hendrik Strobelt, Sebastian Gehrmann, Michael Behrisch, Adam Perer, Hanspeter Pfister, and Alexander M Rush. 2018. Seq2seq-vis: A visual debugging tool for sequence-to-sequence models. *IEEE transactions on visualization and computer graphics*, 25(1):353–363.

[672] Simone Stumpf, Vidya Rajaram, Lida Li, Margaret Burnett, Thomas Dietterich, Erin Sullivan, Russell Drummond, and Jonathan Herlocker. 2007. Toward harnessing user feedback for machine learning. In *Proceedings of the 12th international conference on Intelligent user interfaces*, pages 82–91.

[673] Simone Stumpf, Vidya Rajaram, Lida Li, Weng-Keen Wong, Margaret Burnett, Thomas Dietterich, Erin Sullivan, and Jonathan Herlocker. 2009. Interacting meaningfully with machine learning systems: Three experiments. *International Journal of Human-Computer Studies*, 67(8):639–662.

[674] Theodore R Sumers, Mark K Ho, Robert D Hawkins, Karthik Narasimhan, and Thomas L Griffiths. 2021. Learning rewards from linguistic feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6002–6010.

[675] AI Safety Summit. 2023. The bletchley declaration by countries attending the ai safety summit, 1-2 november 2023. https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023.

[676] Chuangchuang Sun, Macheng Shen, and Jonathan P How. 2020. Scaling up multiagent reinforcement learning for robotic systems: Learn an adaptive sparse communication graph. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 11755–11762. IEEE.

[677] Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. 2024. Principle-driven self-alignment of language models from scratch with minimal human supervision. *Advances in Neural Information Processing Systems*, 36.

[678] Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z Leibo, Karl Tuyls, et al. 2018. Value-decomposition networks for cooperative multi-agent learning based on team reward. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 2085–2087.

[679] Simon Suo, Sebastian Regalado, Sergio Casas, and Raquel Urtasun. 2021. Trafficsim: Learning to simulate realistic multi-agent behaviors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10400–10409.

[680] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.

[681] Justin Svegliato, Samer B Nashed, and Shlomo Zilberstein. 2021. Ethically compliant sequential decision making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35(13), pages 11657–11665.

[682] Shea Swaugerarchive. 2020. Software that monitors students during tests perpetuates inequality and violates their privacy. https://www.technologyreview.com/2020/08/07/1006132/software-algorithms-proctoring-online-tests-ai-ethics.

[683] Umar Syed, Michael Bowling, and Robert E Schapire. 2008. Apprenticeship learning using linear programming. In *Proceedings of the 25th international conference on Machine learning*, pages 1032–1039.

[684] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

[685] Jonas Tallberg, Eva Erman, Markus Furendal, Johannes Geith, Mark Klamberg, and Magnus Lundgren. 2023. The global governance of artificial intelligence: Next steps for empirical and normative research. *arXiv preprint arXiv:2305.11528*.

[686] Kai Liang Tan, Yasaman Esfandiari, Xian Yeow Lee, Soumik Sarkar, et al. 2020. Robustifying reinforcement learning agents via action space adversarial training. In *2020 American control conference (ACC)*, pages 3959–3964. IEEE.

[687] Ming Tan. 1993. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the tenth international conference on machine learning*, pages 330–337.

[688] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.

[689] Annalisa T Taylor, Thomas A Berrueta, and Todd D Murphey. 2021. Active learning in robotics: A review of control principles. *Mechatronics*, 77:102576.

[690] The White House. 2023. Fact sheet: Biden-harris administration secures voluntary commitments from leading artificial intelligence companies to manage the risks posed by ai. https://www.whitehouse.gov/b riefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-admin istration-secures-voluntary-commitments-from-leading-artificial-intellige nce-companies-to-manage-the-risks-posed-by-ai.

[691] Andrea L Thomaz and Cynthia Breazeal. 2008. Teachable robots: Understanding human teaching behavior to build more effective robot learners. *Artificial Intelligence*, 172(6-7):716–737.

[692] Sunil Thulasidasan, Sushil Thapa, Sayera Dhaubhadel, Gopinath Chennupati, Tanmoy Bhattacharya, and Jeff Bilmes. 2021. An effective baseline for robustness to distributional shift. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 278–285. IEEE.

[693] Jeremy Tien, Jerry Zhi-Yang He, Zackory Erickson, Anca Dragan, and Daniel S Brown. 2022. Causal confusion and reward misidentification in preference-based reward learning. In *The Eleventh International Conference on Learning Representations*.

[694] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. 2017. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 23–30. IEEE.

[695] Suzanne Tolmeijer, Markus Kneer, Cristina Sarasua, Markus Christen, and Abraham Bernstein. 2020. Implementations in machine ethics: A survey. *ACM Computing Surveys (CSUR)*, 53(6):1–38.

[696] Faraz Torabi, Garrett Warnell, and Peter Stone. 2018. Behavioral cloning from observation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4950–4957.

[697] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

[698] Robert Trager, Ben Harack, Anka Reuel, Allison Carnegie, Lennart Heim, Lewis Ho, Sarah Kreps, Ranjit Lall, Owen Larter, Seán Ó hÉigeartaigh, et al. 2023. International governance of civilian ai: A jurisdictional certification approach. *arXiv preprint arXiv:2308.15514*.

[699] Johannes Treutlein, Michael Dennis, Caspar Oesterheld, and Jakob Foerster. 2021. A new formalism, method and open issues for zero-shot coordination. In *International Conference on Machine Learning*, pages 10413–10423. PMLR.

[700] Grigorios Tsoumakas and Ioannis Katakis. 2007. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13.

[701] Alexey Turchin and David Denkenberger. 2020. Classification of global catastrophic risks connected with artificial intelligence. *Ai & Society*, 35(1):147–163.

References

[702] Alex Turner. 2022. Inner and outer alignment decompose one hard problem into two extremely hard problems. https://www.alignmentforum.org/posts/gHefoxiznGfsbiAu9/inner-and-outer-alignment-decompose-one-hard-problem-into.

[703] Alex Turner, Neale Ratzlaff, and Prasad Tadepalli. 2020. Avoiding side effects in complex environments. *Advances in Neural Information Processing Systems*, 33:21406–21415.

[704] Alex Turner, Logan Smith, Rohin Shah, Andrew Critch, and Prasad Tadepalli. 2021. Optimal policies tend to seek power. *Advances in Neural Information Processing Systems*, 34:23063–23074.

[705] Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2024. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36.

[706] UNESCO. 2021. Recommendation on the ethics of artificial intelligence. https://unesdoc.unesco.org/ark:/48223/pf0000381137.

[707] UniteAI. 2023. What is ai capability control & why does it matter? https://www.unite.ai/what-is-ai-capability-control-why-does-it-matter.

[708] Fabio Urbina, Filippa Lentzos, Cédric Invernizzi, and Sean Ekins. 2022. Dual use of artificial-intelligence-powered drug discovery. *Nature Machine Intelligence*, 4(3):189–191.

[709] Paul AM Van Lange, Ellen De Bruin, Wilma Otten, and Jeffrey A Joireman. 1997. Development of prosocial, individualistic, and competitive orientations: theory and preliminary evidence. *Journal of personality and social psychology*, 73(4):733.

[710] Vladimir Vapnik. 1991. Principles of risk minimization for learning theory. *Advances in Neural Information Processing Systems*, 4.

[711] Shikhar Vashishth, Shyam Upadhyay, Gaurav Singh Tomar, and Manaal Faruqui. 2019. Attention interpretability across nlp tasks.

[712] Ajit Kumar Verma, Srividya Ajit, Durga Rao Karanki, et al. 2010. *Reliability and safety engineering*, volume 43. Springer.

[713] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *Proceedings of the international workshop on software fairness*, pages 1–7.

[714] Krakovna Victoria, Uesato Jonathan, Mikulik Vladimir, Rahtz Matthew, Everitt Tom, Kumar Ramana, Kenton Zac, Leike Jan, and Legg Shane. 2020. Specification gaming: the flip side of ai ingenuity. https://www.deepmind.com/blog/specification-gaming-the-flip-side-of-ai-ingenuity.

[715] Jesse Vig. 2019. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42.

[716] Ricardo Vinuesa, Hossein Azizpour, Iolanda Leite, Madeline Balaam, Virginia Dignum, Sami Domisch, Anna Felländer, Simone Daniela Langhans, Max Tegmark, and Francesco Fuso Nerini. 2020. The role of artificial intelligence in achieving the sustainable development goals. *Nature Communications*, 11(1):1–10.

[717] Georg Henrik Von Wright. 1951. Deontic logic. *Mind*, 60(237):1–15.

[718] Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing nlp. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

[719] Dilin Wang, Chengyue Gong, and Qiang Liu. 2019a. Improving neural language modeling via adversarial training. In *International Conference on Machine Learning*, pages 6555–6565. PMLR.

[720] Fulton Wang and Cynthia Rudin. 2015. Falling rule lists. In *Artificial intelligence and statistics*, pages 1013–1022. PMLR.

[721] Jiaqi Wang, Huafeng Liu, Xinyue Wang, and Liping Jing. 2021. Interpretable image recognition by constructing transparent embedding space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 895–904.

[722] Kaimeng Wang, Yu Zhao, and Ichiro Sakuma. 2023a. Learning robotic insertion tasks from human demonstration. *IEEE Robotics and Automation Letters*.

References

[723] Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2022. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. In *The Eleventh International Conference on Learning Representations*.

[724] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2023b. A survey on large language model based autonomous agents. *arXiv preprint arXiv:2308.11432*.

[725] Rui Wang, Joel Lehman, Jeff Clune, and Kenneth O Stanley. 2019b. Poet: open-ended coevolution of environments and their optimized solutions. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 142–151.

[726] Woodrow Z Wang and Mark Beliaev. 2021. Emergent prosociality in multi-agent games through gifting. In *30th International Joint Conference on Artificial Intelligence (IJCAI)*.

[727] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023c. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

[728] Zhenyi Wang, Xiaoyang Wang, Bang An, Dong Yu, and Changyou Chen. 2020. Towards faithful neural table-to-text generation with content-matching constraints. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1072–1086.

[729] Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the gap: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.

[730] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2024. Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems*, 36.

[731] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

[732] Jörgen W Weibull. 1997. *Evolutionary game theory*. MIT press.

[733] Laura Weidinger, Kevin R McKee, Richard Everett, Saffron Huang, Tina O Zhu, Martin J Chadwick, Christopher Summerfield, and Iason Gabriel. 2023. Using the veil of ignorance to align ai systems with principles of justice. *Proceedings of the National Academy of Sciences*, 120(18):e2213709120.

[734] Lilian Weng. 2023a. Adversarial attacks on llms. https://lilianweng.github.io/posts/2023-10-25-adv-attack-llm.

[735] Lilian Weng. 2023b. Llm powered autonomous agents. https://lilianweng.github.io/posts/2023-06-23-agent.

[736] Darrell M West. 2018. *The future of work: Robots, AI, and automation*. Brookings Institution Press.

[737] Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart Van Merriënboer, Armand Joulin, and Tomas Mikolov. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*.

[738] White House. 2023. Ensuring safe, secure, and trustworthy ai. https://www.whitehouse.gov/wp-content/uploads/2023/07/Ensuring-Safe-Secure-and-Trustworthy-AI.pdf.

[739] Erik Wijmans, Manolis Savva, Irfan Essa, Stefan Lee, Ari S. Morcos, and Dhruv Batra. 2023. Emergence of maps in the memories of blind navigation agents. In *The Eleventh International Conference on Learning Representations*.

[740] wikipedia. 2023. Existential risk from artificial general intelligence. https://en.wikipedia.org/wiki/Existential_risk_from_artificial_general_intelligence.

[741] Claus O Wilke, Jia Lan Wang, Charles Ofria, Richard E Lenski, and Christoph Adami. 2001. Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature*, 412(6844):331–333.

[742] Alan F Winfield, Katina Michael, Jeremy Pitt, and Vanessa Evers. 2019. Machine ethics: The design and governance of ethical ai and autonomous systems [scanning the issue]. *Proceedings of the IEEE*, 107(3):509–517.

[743] Christian Wirth, Riad Akrour, Gerhard Neumann, Johannes Fürnkranz, et al. 2017. A survey of preference-based reinforcement learning methods. *Journal of Machine Learning Research*, 18(136):1–46.

[744] Christian Wirth and Johannes Fürnkranz. 2013. Preference-based reinforcement learning: A preliminary survey. In *Proceedings of the ECML/PKDD-13 Workshop on Reinforcement Learning from Generalized Feedback: Beyond Numeric Rewards*. Citeseer.

[745] Jeff Wu, Long Ouyang, Daniel M. Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. 2021. Recursively summarizing books with human feedback.

[746] Yueh-Hua Wu and Shou-De Lin. 2018. A low-cost ethics shaping approach for designing reinforcement learning agents. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

[747] Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2024. Fine-grained human feedback gives better rewards for language model training. *Advances in Neural Information Processing Systems*, 36.

[748] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*, pages 1391–1399.

[749] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. 2023. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*.

[750] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. 2018a. Fairgan: Fairness-aware generative adversarial networks. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 570–575. IEEE.

[751] Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2020. Recipes for safety in open-domain chatbots. *arXiv preprint arXiv:2010.07079*.

[752] Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2021. Bot-adversarial dialogue for safe conversational agents. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2950–2968.

[753] Jingyi Xu, Zilu Zhang, Tal Friedman, Yitao Liang, and Guy Broeck. 2018b. A semantic loss function for deep learning with symbolic knowledge. In *International conference on machine learning*, pages 5502–5511. PMLR.

[754] Ning Xu, Brian Price, Scott Cohen, Jimei Yang, and Thomas S Huang. 2016. Deep interactive object selection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 373–381.

[755] Mengjiao Yang, Sergey Levine, and Ofir Nachum. 2021. Trail: Near-optimal imitation learning with suboptimal data. In *International Conference on Learning Representations*.

[756] Sherry Yang, Ofir Nachum, Yilun Du, Jason Wei, Pieter Abbeel, and Dale Schuurmans. 2023. Foundation models for decision making: Problems, methods, and opportunities. *arXiv preprint arXiv:2303.04129*.

[757] Tianhe Yu Yevgen Chebotar. 2023. Rt-2: New model translates vision and language into action. https://www.deepmind.com/blog/rt-2-new-model-translates-vision-and-language-into-action.

[758] Zheng-Xin Yong, Cristina Menghini, and Stephen H Bach. 2023. Low-resource languages jailbreak gpt-4. *arXiv preprint arXiv:2310.02446*.

[759] Jin Yong Yoo and Yanjun Qi. 2021. Towards improving adversarial training of nlp models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 945–956.

[760] Chao Yu, Jiming Liu, Shamim Nemati, and Guosheng Yin. 2021. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(1):1–36.

[761] Han Yu, Zhiqi Shen, Chunyan Miao, Cyril Leung, Victor R Lesser, and Qiang Yang. 2018. Building ethics into artificial intelligence. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 5527–5533.

[762] Wenhao Yu, Nimrod Gileadi, Chuyuan Fu, Sean Kirmani, Kuang-Huei Lee, Montserrat Gonzalez Arenas, Hao-Tien Lewis Chiang, Tom Erez, Leonard Hasenclever, Jan Humplik, brian ichter, Ted Xiao, Peng Xu, Andy Zeng, Tingnan Zhang, Nicolas Heess, Dorsa Sadigh, Jie Tan, Yuval Tassa, and Fei Xia. 2023. Language to rewards for robotic skill synthesis. In *7th Annual Conference on Robot Learning*.

[763] Yang Yu. 2018. Towards sample efficient reinforcement learning. In *IJCAI*, pages 5739–5743.

[764] Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. 2024. Rrhf: Rank responses to align language models with human feedback. *Advances in Neural Information Processing Systems*, 36.

[765] Luyao Yuan, Xiaofeng Gao, Zilong Zheng, Mark Edmonds, Ying Nian Wu, Federico Rossano, Hongjing Lu, Yixin Zhu, and Song-Chun Zhu. 2022. In situ bidirectional human-robot value alignment. *Science Robotics*, 7(68):eabm4183.

[766] E Yudkowsky. 2018. Challenges to christiano's capability amplification proposal. *LessWrong*.

[767] Man-Ching Yuen, Irwin King, and Kwong-Sak Leung. 2011. A survey of crowdsourcing systems. In *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*, pages 766–773. IEEE.

[768] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 1415–1420. Association for Computational Linguistics.

[769] Yuan Zang, Fanchao Qi, Chenghao Yang, Zhiyuan Liu, Meng Zhang, Qun Liu, and Maosong Sun. 2020. Word-level textual adversarial attacking as combinatorial optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6066–6080.

[770] Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 818–833. Springer.

[771] Rowan Zellers. 2019. Why we released grover. https://thegradient.pub/why-we-released-grover.

[772] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018a. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340.

[773] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2018b. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*.

[774] Quan-shi Zhang and Song-Chun Zhu. 2018. Visual interpretability for deep learning: a survey. *Frontiers of Information Technology & Electronic Engineering*, 19(1):27–39.

[775] Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu. 2018c. Interpretable convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8827–8836.

[776] Shiyin Zhang, Jun Hao Liew, Yunchao Wei, Shikui Wei, and Yao Zhao. 2020a. Interactive object segmentation with inside-outside guidance. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12234–12244.

[777] Tianhao Zhang, Zoe McCarthy, Owen Jow, Dennis Lee, Xi Chen, Ken Goldberg, and Pieter Abbeel. 2018d. Deep imitation learning for complex manipulation tasks from virtual reality teleoperation. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5628–5635. IEEE.

[778] Tianjun Zhang, Fangchen Liu, Justin Wong, Pieter Abbeel, and Joseph E. Gonzalez. 2023a. The wisdom of hindsight makes language models better instruction followers. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 41414–41428. PMLR.

[779] Wenjia Zhang, Haoran Xu, Haoyi Niu, Peng Cheng, Ming Li, Heming Zhang, Guyue Zhou, and Xianyuan Zhan. 2023b. Discriminator-guided model-based offline imitation learning. In *Conference on Robot Learning*, pages 1266–1276. PMLR.

[780] Yuan Zhang, Xiaoran Xu, Hanning Zhou, and Yan Zhang. 2020b. Distilling structured knowledge into embeddings for explainable and accurate recommendation. In *Proceedings of the 13th international conference on web search and data mining*, pages 735–743.

[781] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023c. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

[782] Zhexin Zhang, Jiale Cheng, Hao Sun, Jiawen Deng, Fei Mi, Yasheng Wang, Lifeng Shang, and Minlie Huang. 2022. Constructing highly inductive contexts for dialogue safety through controllable reverse generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3684–3697.

[783] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20.

[784] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

[785] Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Man Cheung, and Min Lin. 2024. On evaluating adversarial robustness of large vision-language models. *Advances in Neural Information Processing Systems*, 36.

[786] Stephan Zheng, Yang Song, Thomas Leung, and Ian Goodfellow. 2016. Improving the robustness of deep neural networks via stability training. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 4480–4488.

[787] Bolei Zhou, Yiyou Sun, David Bau, and Antonio Torralba. 2018. Revisiting the importance of individual units in cnns via ablation. *arXiv preprint arXiv:1806.02891*.

[788] Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2024. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36.

[789] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. 2022. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[790] Li Zhou and Kevin Small. 2021. Inverse reinforcement learning with natural language goals. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11116–11124.

[791] Zhi-Hua Zhou. 2021. *Machine learning*. Springer Nature.

[792] Henry Zhu, Justin Yu, Abhishek Gupta, Dhruv Shah, Kristian Hartikainen, Avi Singh, Vikash Kumar, and Sergey Levine. 2019. The ingredients of real world robotic reinforcement learning. In *International Conference on Learning Representations*.

[793] Kaijie Zhu, Jiaao Chen, Jindong Wang, Neil Zhenqiang Gong, Diyi Yang, and Xing Xie. 2023. Dyval: Graph-informed dynamic evaluation of large language models. *arXiv preprint arXiv:2309.17167*.

[794] Simon Zhuang and Dylan Hadfield-Menell. 2020. Consequences of misaligned ai. *Advances in Neural Information Processing Systems*, 33:15763–15773.

[795] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al. 2008. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pages 1433–1438. Chicago, IL, USA.

[796] Daniel Ziegler, Seraphina Nix, Lawrence Chan, Tim Bauman, Peter Schmidt-Nielsen, Tao Lin, Adam Scherlis, Noa Nabeshima, Benjamin Weinstein-Raun, Daniel de Haas, et al. 2022. Adversarial training for high-stakes reliability. *Advances in Neural Information Processing Systems*, 35:9274–9286.

[797] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

[798] Caleb Ziems, Jane Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2022. The moral integrity corpus: A benchmark for ethical dialogue systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3755–3773.

[799] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. 2023a. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

[800] Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. 2023b. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

[801] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. 2024. Segment everything everywhere all at once. *Advances in Neural Information Processing Systems*, 36.