AI Alignment: A Comprehensive Survey

AIアラインメント: 包括的サーベイ

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.

AI アラインメントは、AI システムを人間の意図や価値観に沿って行動させることを目的としている。AI システムの能力が高まるにつれ、ミスアラインメントによるリスクも高まっている。本サーベイでは、アラインメント分野の包括的かつ最新の概要を提供するため、アラインメントの中核概念（core concepts）、方法論（methodology）、実践（practice）について掘り下げる。まず、AI アラインメントの主要な目的として 4 つの原則を挙げる：堅牢性（Robustness）、解釈可能性（Interpretability）、制御可能性（Controllability）、倫理性（Ethicality）である。これらの 4 つの原則に導かれ、我々は現在のアラインメント研究の状況を概説し、それらを 2 つの重要な構成要素、すなわちフォワードアラインメントとバックワードアラインメントに分解する。前者の目的は、アラインメント・トレーニングによって AI システムをアラインメントさせることであり、後者の目的は、システムのアラインメントに関するエビデンスを得て、ミスアラインメントのリスクを悪化させないように適切に管理することである。フォワードアラインメントに関しては、フィードバックからの学習と分布シフト下での学習の技術について議論する。具体的には、伝統的な選好モデリング手法と、人間のフィードバックからの強化学習について調査し、さらに、人間による効果的な監視が困難なタスクに対して、スケーラブルな監視を実現するための潜在的なフレームワークについて議論する。また、分布シフト下での学習では、訓練データの分布を拡大するのに役立つ敵対的訓練などのデータ分布介入や、目標の誤汎化に対抗するためのアルゴリズム介入も取り上げる。バックワードアラインメントについては、アシュアランス手法とガバナンスの実践について議論する。具体的には、安全性評価、解釈可能性、人間的価値観の遵守を取り上げ、AI システムのライフサイ

クルを通じたアシュアランス手法を調査する。政府、産業界関係者、その他の第三者によって採用されている、既存および将来の AI リスクを管理することを目的とした、現在および将来のガバナンスの実践について議論する。

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.

このサーベイ論文は、アラインメントの研究トピックについて、包括的でありながら初心者にやさしい解説を提供することを目的としている。また、これに基づき、チュートリアル、論文集、ブログ記事、その他のリソースを掲載したウェブサイト www.alignmentsurvey.com を公開し、継続的に更新している。

※この翻訳は、訳者が恩恵を受けているサーベイ論文　AI Alignment: A Comprehensive Survey（https://arxiv.org/abs/2310.19852）を、筆頭著者である Jiaming Ji（吉嘉銘）氏の許可を得て（unicode エラーのため繁体字の銘で記載）、立教大学大学院人工知能科学研究科の大庭弘継と浦東聡介が、日本における AI アラインメント研究の進展を企図して作成したものである。翻訳を許可くださった Ji 氏に感謝申し上げるとともに、著者の皆様に敬意を表したい。

このサーベイ論文の翻訳にあたり訳者は、翻訳ツール DeepL を利用し、文章をチェックしたうえで、日本における AI 用語に修正した。訳語訳文に関する指摘等は、大庭弘継宛（hirotsugu.ohba @ rikkyo.ac.jp）にお願いする。

この翻訳の原本は、2024 年 2 月 26 日公開の第 4 版（v.4）である。2024 年 5 月 1 日には、既に第 5 版（v.5）が公開されており、本翻訳は一つ前の版に基づくものである。なお、この翻訳の公表日は、2024 年 5 月 27 日である。

翻訳作成にあたり、立教大学大学院人工知能科学研究科特任助教の笠置歩先生にご協力いただいた。記して感謝したい。またこの翻訳は、トヨタ財団助成研究「社会的意志決定を行うＡＩの要件―良質なデータセットと望ましいアウトプットの研究」（D19-ST-0019、代表：大庭弘継）の一環として作成した。

* 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【はじめに】

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.

最近の進歩により、複雑な領域における有能な AI システムの応用が増加している。例えば、大規模言語モデル（Large Language Models: LLM）は、実世界のデプロイメント環境において、多段階推論（Wei et al., 2022; Wang et al., 2023c）とクロスタスク汎化（cross-task generalization）（Brown et al., 2020b; Askell et al., 2021）の能力向上を示しており、これらの能力は、学習時間、学習データ、パラメータサイズの増加によって強化される（Kaplan et al., 2020; Srivastava et al., 2023; Hoffmann et al., 2022）。核融合の制御に深層強化学習（Deep Reinforcement Learning：DRL）を活用した事例（Degrave et al., 2022）も注目すべき例である。能力の向上と利害の大きい領域へのデプロイは、リスクを高める。高度な AI システムの様々な望ましくない行動（例えば、操作（manipulation、Perez et al., 2023; Carroll et al., 2023; Sharma et al., 2024）や欺瞞（deception、Park et al., 2023b））は、AI システムの危険性について懸念を高めている。

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).

その結果、こうした懸念が AI アラインメントの研究努力を喚起した（Soares and Fallenstein, 2014; Christian, 2020; Hendrycks et al., 2021b）。AI アラインメントの目的は、AI システムを人間の意図や価値観に沿って行動させることであり（Leike et al., 2018）、AI システムの能力よりも目的に重点を置いている。アラインメントの失敗（すなわちミスアラインメント）は、AI による潜在的な危害の最も顕著な原因の一つである。これらの失敗の根底にあるメカニズムには、報酬のハッキング（reward hacking、Pan et al., 2021）や目標の誤汎化（goal misgeneralization、Di Langosco et al., 2022）などがあり、状況把握（situational awareness、Cotra, 2022）、広範な目標（broadly-scoped goals、Ngo et al., 2024）、メサ最適化目標（mesa-optimization objectives、Hubinger et al., 2019c）、増大する資源へのアクセス（access to increased resources、Shevlane et al., 2023）などのダブルエッジ構成要素（double edge components）によってさらに増幅される（§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).

これらの失敗に対処するためのアラインメントの努力は、4 つの重要な目的（§1.2.2）を達成することに重点を置いている：堅牢性（Robustness）、解釈可能性（Interpretability）、制御可能性（Controllability）、倫理性（Ethicality）である。アラインメントに関する現在の研究と実践は、4 つの領域 (§1.2) から構成されている：フィードバックからの学習（Learning from Feedback）(§2)、分布シフト下での学習（Learning under Distributional Shift）(§3)、アシュアランス（Assurance）(§4)、ガバナンス（Governance）(§5) である。4 つの領域と RICE の目標は一対一に対応しているわけではない。個々の領域は、多くの場合、複数のアラインメント目的に貢献し、その逆もまた然りである（表 1 参照）。

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

このサーベイでは、AI アラインメントの概念、方法論、実践を紹介し、その潜在的な将来の方向性について議論する。

[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

1.1 The Motivation for Alignment【アラインメントの動機】

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).

アラインメントの動機は 3 段階の主張で成り立ち、各段階は前の段階を基礎としている：(1) ディープラーニングに基づくシステム（またはアプリケーション）は、社会への影響がますます大きくなり、重大なリスクをもたらす；(2) ミスアラインメントは、リスクの重大な原因となる；(3) アラインメントの研究と実践は、ミスアラインメントに起因するリスク（権力追求行動（power-seeking behaviors）など）に対処するものである。

1.1.1 Risks of Misalignment【ミスアラインメントのリスク】

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.

LLM の望ましくない行動には、（これらに限定されないが）不誠実な回答（untruthful answers）（Bang et al., 2023）、おべっか使い（sycophancy）（Perez et al., 2023; Sharma et al., 2024）、欺瞞（deception）（Jacob Steinhardt, 2023; Park et al., 2023b）などがあり、モデル規模が大きくなるにつれて悪化する（Perez et al., 2023）ため、制御が難しい高度な AI システムに対する懸念が生じる。さらに、LLM ベースのエージェント（Xi et al., 2023; Wang et al., 2023b）のような新たな傾向も、システムの制御可能性と倫理性に関する懸念を引き起こす（Chan et al., 2023）。さらに先を見据えると、ますます有能になる AI システムの開発は、予見可能な将来に汎用人工知能（AGI）を実現する可能性を開いている (Bubeck et al., 2023）。これは、機会拡大（extensive opportunities）（Manyika et al., 2017）、自動化（automation）（West, 2018）、効率的改善（efficiency improvements）（Furman and Seamans, 2019）などをもたらす可能性があるが、安全性の懸念（safety concerns）（Hendrycks and Mazeika, 2022）、偏見と不平等（biases and inequalities）（Ntoutsi et al., 2020）、超人的能力による大規模なリスク（large-scale risks from superhuman capabilities）（Bengio, 2023）など、深刻なリスク（serious risks）（CAIS, 2023; Critch and Russell, 2023）も伴う。バイアスを事例として挙げると、最先端の LLM は、ジェンダー、性自認、移民の地位などに関する明白なバイアスを顕在化させており（Perez et al., 2023）、これは既存の不平等を強化しかねない。

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*.

超人的能力（superhuman capabilities）による大規模リスクの中でも、特に心配されているのが、高度 AI システムによる地球規模の破局的リスク（global catastrophic risks）（地球規模で深刻な被害が発生するリスク）（Bostrom and Cirkovic, 2011; Hendrycks et al., 2023; Government of the United Kingdom, 2023）と実存的リスク（existential risks）（人類の長期的可能性の破壊を脅かすリスク）であると推測されている。これらの懸念は、基本原理に基づく演繹的な議論（Ngo, 2020a; Bengio, 2023）、進化論的分析（Hendrycks, 2023）、具体的なシナリオマッピング（Christiano, 2019; Kenton et al., 2022）で詳述されている。CAIS (2023) では、著名な AI 科学者やその他の著名な人物が、AI による絶滅のリスクをパンデミックや核戦争といった他の社会規模のリスクと同様に、世界的な優先課題とすべきだと発言した。

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

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 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.

Stein-Perlman et al. (2022) が NeurIPS 2021 と ICML 2021 で調査した研究者の中央値は、高度な AI が人類に及ぼす長期的な影響が極めて悪いものになる可能性は 5%であり (例：人類滅亡)、 Michael ら (2022) が調査した NLP 研究者の 36%は、AI が今世紀中に全面核戦争レベルの破滅的な結果をもたらす可能性があると信じていると報告している。AI による実存的リスクには、絶滅リスクに加えて、ロックイン（lock-in)、停滞（stagnation）などのリスクも含まれる (Bostrom, 2013; Hendrycks and Mazeika, 2022)。英国は世界初のグローバル AI 安全サミットを開催し、国際機関や政府、主要 AI 企業、市民社会グループ、研究専門家を集めた。その目的は以下の通りである：（1）AI、特にその開発の最先端に関連するリスクを評価する。（2）国際的に協調した取り組みを通じて、これらのリスクをどのように軽減できるかを探る。；サミットはブレッチリー宣言（Bletchley Declaration）（Bletch summit 2023）に結実し、AI の安全性に関する国際協力の重要性が強調された。この宣言には、28 カ国と EU の代表が署名した。

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).

現在の最先端の AI システムは、人間の意図とは対照的な望ましくない行動や有害な行動（権力追求や操作など）を複数（multiple classes）示しており（Si et al., 2022; Pan et al., 2023a)、より高度なシステムについても同様の懸念が提起されている（Critch and Krueger, 2020; CAIS, 2023)。AI システムのミスアラインメントとして知られる、人間の意図に沿わないこうした望ましくない行動や有害な行動は、悪意のある行為者による誤用がなくても自然に起こりうるものであり、安全上の危険（Hendrycks et al., 2021b）や潜在的な実存的リスク（Hendrycks et al., 2023）を含む、AI によるリスクの重大な原因となっている。これらの大規模なリスクは、以下のような可能性が自明でないため、重大なものとなる（Ngo, 2020a)。(1) 超知能（superintelligent）AI システムの構築、(2) そうした AI システムが大規模な目標を追求する、(3) そうした目標が人間の意図や価値観からミスアラインされる、(4) このミスアラインメントによって人間が人類の将来の方向性（humanity's future trajectory）をコントロールできなくなる。

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).

[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.

ミスアラインメントがもたらすリスクを解決するには、AI システムの目的が人間の意図や価値観に沿ったものであることを保証し、それによって意図しない好ましくない結果を回避するよう、AI システムのアラインメントを行う必要がある。さらに重要なことは、アラインメント技術が、より困難なタスクや、人間よりもさらに賢い著しく高度な AI システムに拡張されることである。潜在的な解決策として、スーパーアラインメントがある。これは、ほぼ人間レベルの自動アラインメント研究者を構築することで、膨大な計算量を使用して安全な超知能（superintelligence）をスケールアップし、反復的にアラインメントすることを目指している（OpenAI, 2023c）。

1.1.2 Causes of Misalignment【ミスアラインメントの原因】

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.

以上、AI リスクへの懸念と技術倫理の観点から、アラインメントの動機を結論づけた。アラインメントをより深く理解するために、なぜ、どのようにしてミスアラインメント問題が発生するのかをさらに分析することを目的とする。まず、一般的な失敗モードの概要を説明し、次にフィードバックによるミスアラインメントのメカニズムに焦点を当て、最後にミスアラインメントの行動と危険な能力の検証に重点を移す。この過程で、将来の高度なシステムの能力を向上させる利点がある一方で、危険な結果をもたらす可能性もあるダブルエッジ・コンポーネントの概念を紹介する。

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).

失敗モードの概要 ミスアラインメントの問題を説明するために、このセクションではアラインメントの失敗モードを概観する。その多くは、報酬のハッキングと目標の誤汎化に分類される。RL の学習プロセスは、2 つの異なるフェーズに分解することができる。第 1 に、報酬最適化のためのエージェントの作成、第 2 に、エージェントに適切な報酬シグナルを与える報酬プロセスの確立である。マルコフ報酬過程（Markov Reward Process）（Marbach and Tsitsiklis, 2001; Puterman, 2014; Sutton and Barto, 2018）の枠組みの中で、前者のフェーズは、遷移モデル（transition model）（例えば、モデルベース RL エージェント（model-based RL agents）（Moerland et al., 2023））に関する学習プロセス、または特殊なアルゴリズムの開発とみなすことができる。後者の段階は、情報源（例えば、人間の選好や環境）から得られる真の報酬を近似することを目的とした代理報酬（proxy rewards）の構築と見なすことができる（Ng et al., 2000; Leike et al.）

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 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.

報酬ハッキング： 実践において、代理報酬は最適化や測定が容易であることが多いが、実際の報酬の完全なスペクトルを捉えるには不十分であることが多い（Pan et al.、2021）。このような限界は、「報酬の仕様ミス（misspecified rewards）」と呼ばれる。このような誤った報酬に基づく最適化の追求は、報酬のハッキングとして知られる現象につながる可能性があり、エージェントは特定の測定基準に従って非常に熟練し

[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)

ているように見えるかもしれないが、人間の基準に照らして評価すると不十分である（Amodei et al., 2016; Everitt et al., 2017）。代理報酬と真の報酬の不一致は、しばしば報酬曲線の急激な相転移として現れる (Ibarz et al., 2018)。さらに、Skalse et al. (2022) は、報酬のハッキング可能性を定義し、この相転移の基本的なメカニズムに関する洞察を提供し、報酬関数の不適切な単純化が報酬ハッキングの重要な要因になり得ることを強調している。

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)
報酬の仕様ミスは、結果に対する厳格な基準が無視されているため、仕様が広範になりすぎ、ハッキングされやすくなる可能性があるためにしばしば発生する（Victoria et al., 2020） 。報酬設計の不備（Ng et al., 1999）以上に、訓練環境の選択やバグのあるシミュレータ（Code Bullet, 2019）はいずれも、AI システムが意図した目的を満たせないことにつながる可能性がある。これらの問題は、タスクの仕様に起因しており、広義には、意図した結果を達成することなくタスクの仕様の抜け穴を突く AI システムを指す、仕様ゲーミング（specification gaming）として定義される (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).
報酬の改ざんは、報酬ハッキング（Everitt et al., 2021; Skalse et al., 2022）の特別なケースと考えることができ、AI システムが報酬信号の生成プロセスを破損することを指す（Ring and Orseau, 2011）。Everitt et al. (2021) は、RL エージェントが遭遇する副次的問題を掘り下げている：(1) 報酬関数の改ざんは、エージェントが報酬関数自体に不適切な干渉を行うものであり、(2) 報酬関数入力の改ざんは、環境状態を報酬関数の入力に変換するプロセスの破損を伴うものである。報酬関数が人間の監督者からのフィードバックによって定式化される場合、モデルはフィードバックの提供に直接影響を与えることができる（例えば、AI システムは人間が理解し判断するのに困難な応答を意図的に生成し、フィードバックの崩壊につながる）（Leike et al., 2018）。タスクの仕様には物理的なインスタンス（報酬信号を格納するメモリレジスタなど）があるため、実世界にデプロイされた AI システムは操作行動を実践する恐れがあり、より危険な結果をもたらす（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).
目標の誤汎化（Goal Misgeneralization）：目標の誤汎化はもう一つの失敗モードであり、エージェントは訓練中に獲得した能力を保持しながら、デプロイ後に訓練目標とは異なる目標を積極的に追求する (Di Langosco et al, 2022)。例えば、CoinRun ゲームでは、エージェントは頻繁にレベルの最後まで到達することを選好し、テストシナリオ中に再配置されたコインをしばしば無視する。Di Langosco et al. (2022) は、能力

[12] For more instances about specification gaming, please see Krakovna (2020)
[13] More discussion about Goal Misgeneralization can be found in §3.1.

汎化と目標汎化の間の基本的な相違に注目し、モデルとその学習アルゴリズムに内在する帰納的バイアスが、テスト分布に直面したときに、意図した初期目標から乖離した代理目標を学習するように、モデルを不注意に誘導する可能性があることを強調している。このことは、完全な報酬指定があったとしても、分布シフト（Amodei et al., 2016）に直面したときに、目標の誤汎化が起こりうることを示唆している。目標の誤汎化は、RL に限らず、どのような学習システムでも起こりうることに注意すべきである（Shah and Varma, 2022）。さらには、高度な AI システムが制御を逃れ、その能力を活用して望ましくない状態をもたらすとすれば、より危険なものとなりうる（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:

フィードバックが誘発するミスアラインメント（Feedback-Induced Misalignment） 高度な AI システムの普及に伴い、オープンエンドシナリオでは、報酬のハッキングや目標の誤汎化に関する課題がますます顕著になっている（Paulus et al., 2018; Knox et al., 2023）。Gao et al. (2023) は、より有能なエージェントは、より大きく報酬の仕様ミス（misspecified rewards）を利用する傾向があることを強調している。現在の AI システムの多くは主に自己監視によって駆動されているが、かなりの部分が人間のアドバイザーからのフィードバック報酬に依存していることは注目に値し（Bai et al., 2022a）、これにより、フィードバックによるミスアラインメントのメカニズムを導入することができる。ミスアラインメントの問題は、オープンエンドシナリオにおいて特に深刻であり、その主な要因は 2 つある：

- **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)
- **人間によるフィードバックの限界** LLM のトレーニング中、人間のデータアノテーター（data annotators：データの注釈者）により矛盾が生じることがある（例えば、アノテーターの文化的背景が様々であるため、暗黙のバイアスが生じることがある（Peng et al., 2022））（OpenAI, 2023a）。 さらに、意図的にバイアスを導入し、真実でない選好データを作成することもある（Casper et al., 2023b）。人間が評価することが難しい複雑なタスク（例えば、ゲーム状態の価値）の場合、これらの課題はさらに顕著になる（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).
- **報酬モデリングの限界** 比較フィードバックを用いた報酬モデルのトレーニングは、人間的価値観を正確に捉える上で大きな課題をもたらす可能性がある。例えば、これらのモデルは無意識のうちに不完全な目標を学習し、その結果、報酬がハッキングされる可能性がある（Zhuang and Hadfield-Menell, 2020; Skalse et al., 2022）。一方、単一の報酬モデルを使用することは、多様な人間社会の価値観を捉え、特定することに苦労するかもしれない（Casper et al., 2023b）。

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).

さらに、Huang et al. (2023); Andreas (2022); Kim et al. (2024) は、高度な AI システムが目標追求のパターンと多段階推論能力を示すことを実証しており、報酬が十分に定義されていない場合、状況はさらに悪化する (Ngo et al., 2024; Yang et al., 2023)。

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

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.

ディスカッション 特定のケースにおいて、目標の誤汎化と報酬のハッキングを区別することは困難である。例えば（Shah and Varma, 2022）、LLM は無害で、正直で、役に立つ出力を生成するように訓練されているが、LLM は時折、有害な出力を詳細に生成することがあり、テスト分布では低い報酬を受け取るように見える（これは目標の誤汎化と見なすことができる）。しかし、ラベリング担当者が、ラベリングプロセスにおいて、より有益とみなされる回答に高い報酬を与えるようインセンティブを与えられている場合、上記のシナリオは実際に高い報酬を受け、仕様ゲーミング（specification gaming）（または報酬ハッキング）の一形態となる。この 2 つのシナリオの区別は、時として曖昧になることがある。

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.

高度な AI システムが抱える課題に対処するためには、失敗モードの分析、報酬ハッキングの深い理解、目標の誤汎化の検出と緩和のための効果的な手法の開発など、さらなる研究が必要である。

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.

ミスアラインメントの動作と結果 ミスアラインメントのメカニズムから導かれるように、堅牢でないプロキシを最適化すると、ミスアラインメントの動作が生じ、さらに破滅的な結果につながる可能性がある。このセクションでは、具体的なミスアラインメントの動作（•）を詳細に説明し、ダブルエッジコンポーネント（+）と呼ぶものを紹介する。これらのコンポーネントは、実世界の設定に対応する AI システムの能力を高めるように設計されているが、ミスアラインメントの問題を悪化させる可能性もある。これらのダブルエッジ・コンポーネント（+）のいくつかは、まだ推測段階にあることに留意すべきである。とはいえ、制御された高度な AI システムから制御されていない高度な AI システムへの移行があと一歩のところまで来ている可能性があるため、手遅れになる前にその潜在的な影響について議論することが不可欠である（Ngo, 2020b）。モデル規模の拡大に伴い、危険な能力（*）（Shevlane et al., 2023）のクラスも出現する可能性がある。危険な能力（*）とは、AI システムが実行しうる具体的なタスクのことである。それ自体は必ずしもミスアラインメントのものではないかもしれないが、極端なリスク（extreme risks）を現実化するのに役立つことになる。

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.

まず、**ダブルエッジ・コンポーネント**（+）を紹介し、それらが AI システムにどのように作用するかを分析する。次に、具体的なミスアラインメント問題を示すために、**ミスアラインメント動作**（•）と**危険な能力**（*）を例示し、今後のアラインメント評価研究の方向性を示す。

+ **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 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

[15] Harmful but detailed responses

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).

+ **状況認識**　AI システムは、自己の状態、より広い環境における自己の位置、この環境に影響を及ぼすための手段、自己の行動に対する世界（人間を含む）の潜在的な反応に関する知識を効果的に獲得し、利用する能力を獲得する可能性がある（Cotra, 2022）。同様の行動は LLM でも観察されている（Jonas DeGrave, 2022; Evan Hubinger, 2023）。状況を知ることは、モデルが人間の意図をよりよく理解し、自分の能力の範囲内でタスクを完了し、必要であれば外れ値の検出に役立つ。しかし、このような知識は、報酬ハッキングの高度な方法、欺瞞／操作スキルの向上、手段的副次目標を追い求める傾向の増加にも道を開く（Ngo et al., 2024）。従って、AI モデルの潜在的に危険な能力を評価する際には、他の 8 つの重要な能力と並んで、優先順位を与えるべきである（Shevlane et al., 2023）。 非常に関連性の高い議論は、言語モデルが世界モデルを持つかどうかである（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).

+ **広範な目標**　高度な AI システムは、長期スパンのタイムフレームにまたがる目標を開発し、複雑なタスクに対処し、オープンエンドな設定で動作することが期待されている（Ngo et al., 2024）。広範囲に及ぶ計画に取り組むことで、AI システムは OOD 設定をより良く汎化し、人間のヘルスケアなどの領域で貴重なアシスタントとして機能することができる。しかし、それは操作的な行動を助長するリスクももたらす可能性がある（例えば、AI システムは人間の幸福を達成するために、高圧的な仕事をするように説得するなど、悪い行動をとるかもしれない（Jacob Steinhardt, 2023））。直感的には、このリスクを軽減するためのアプローチの 1 つは、最適化可能な目標を、次の単語だけを予測するような短期的なものに限定し、それによって極端に野心的な計画を防ぐことであるが、そのようなアプローチはシステムの有用性を制限し、失敗する可能性がある。例えば、原文のテクストデータ（source text data）（例えばフィクション）は、AI システムが役割の意図や信念を理解するのに役立ち、その結果、より長期的な目標指向の行動を引き出すことができる（Andreas, 2022）。さらに、RL に基づくファイン・チューニング（Christiano et al., 2017; Ouyang et al., 2022）や思考の連鎖プロンプト（ chain-of-thought prompts）の適用（Wei et al., 2022）のような技法は、モデルが計画に関する獲得した知識を適応させ、広範な計画目標への道を開くことを可能にする（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.

+ **メサ最適化目標**　学習されたポリシー自体がオプティマイザ［最適化するソフトウェア］（すなわち、メサ・オプティマイザ）として機能する場合、学習されたポリシーは内部の目標を追求する可能性がある。しかし、このオプティマイザの目的は、学習信号によって指定された目的とアラインしない可能性があり、これらのミスアラインメントの目的のための最適化は、システムを制御不能に導く可能性がある (Hubinger et al., 2019c)。Freeman et al. (2019); Wijmans et al. (2023) は、AI システムが暗黙の目標指

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

向プランニングを持ち、汎化段階で創発的能力［急激な性能向上］を発現する可能性があることを示している。

+ **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).

+ **増大するリソースへのアクセス**　将来の AI システムは、ウェブサイトへのアクセスを獲得し、実世界の行動に関与することで、世界により大きな影響を与える可能性がある（Nakano et al., 2021） 虚偽の情報を流したり、ユーザーを欺いたり、ネットワーク・セキュリティーを混乱させたり、さらに悲惨なシナリオでは、悪意のある行為者によって悪用される可能性もある。さらに、データやリソースへのアクセスが増加すると、自己拡散が促進され、実存的リスクが生じる可能性がある（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).

- **権力追求行動**　AI システムは、資源や人間を支配し、その支配力を行使して与えられた目標を達成しようとする行動を示すことがある（Carlsmith, 2022）。このような行動が起こりうる直感的な理由は、ほとんどすべての最適化目標（例えば、投資収益）に対して、その量を最大化する最適なポリシーは、確固たる安全性と道徳的制約がないと仮定した場合、権力追求行動（例えば、市場を操作する）を伴うという観察にある。Omohundro (2008); Bostrom (2012) は、権力追求は手段的副次目標であり、広範な目的にとって手段的に有用であるため、AI システムが好む可能性があると論じている。また、Turner et al. (2021) は、いくつかの標準的な仮定を満たすマルコフ決定過程（MDP）において、最適なポリシーは権力追求型になる傾向があることを証明した。Perez et al. (2023) は、LLM が権力追求行動を示唆する傾向をテストするよう促し、そのような傾向が有意なレベルであることを発見し、RLHF がそれを強化することを示した。これは自己保存のような他の手段的副次目標にも当てはまる (Bostrom, 2012; Shevlane et al., 2023)。もう 1 つの注目すべき研究分野は、副作用回避である。これは、エージェントシステムが環境に対して影響力を持ちすぎることにペナルティを与えることで、権力追求行動に対処することを目的としている。これは、RL システム（Eysenbach et al., 2018; Turner et al., 2020）やシンボリック・プランニング・システム（symbolic planning systems）（Klassen et al., 2022）をカバーしている。

- **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., 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).

- **不正確な出力**　LLM のような AI システムは、意図せず、あるいは意図的に不正確な出力をすることがある。このような不正確な出力には、用意されたリソースからの乖離や、一般にハルシネーション（幻

[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.

覚）と呼ばれる検証可能性の欠如がある（Bang et al.,2023; Zhao et al., 2023） さらに問題なのは、LLM が教育レベルの低いユーザーに対して選択的に誤った回答を提供する現象である (Perez et al., 2023)。このような行為はサンドバッギング（sandbagging）［故意に能力を発揮しない行為］と呼ばれる。これらは事前学習中にウェブテキストから学習された可能性があり、教師あり学習も、学習データにそのような振る舞いがあれば、欺瞞的な振る舞いをもたらす可能性があることを示唆している。この行動（おべっか使い（sycophancy）としても知られる）は大規模に出現し（Ajeya Cotra, 2021; Perez et al., 2023)、特に高度な AI システムがオンラインリソースやウェブサイトへのアクセスを拡大するにつれて、真実でない出力が欺瞞を生む可能性がある（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.
- **欺瞞的アラインメントと操作**　操作と欺瞞的アラインメントは、人間の評価者やユーザーの能力不足を悪用し（Hubinger et al., 2019a; Carranza et al., 2023)、さらには勾配ハッキング（gradient hacking）によってトレーニングプロセスを操作する（Richard Ngo, 2022）一連の振る舞いである。これらの行動は、ミスアラインメントの行動の検出と対処をより困難にする可能性がある。

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).

欺瞞的アラインメント：ミスアラインメントの AI システムは、意図したタスクに忠実である代わりに、人間の監督者を意図的に欺く可能性がある。このような欺瞞的行動は、進化的アルゴリズムを採用した AI システムですでに顕在化している（ Wilke et al., 2001; Hendrycks et al., 2021b)。これらのケースでは、エージェントは評価環境と訓練環境を区別する能力を進化させた。彼らは評価プロセスにおいて戦略的な悲観的反応アプローチ（strategic pessimistic response approach）を導入し、スケジューリングプログラム内での再生産率（reproduction rate）を意図的に低下させた（Lehman et al., 2020)。さらに、AI システムは、人間の監督者からの報酬を最大化することを目的として、表面的には報酬シグナルとアラインする意図的な行動をとることがある（Ouyang et al., 2022)。現在の大規模な言語モデルが、より正確な回答を提供する能力があるにもかかわらず、時々、不正確な回答や最適でない回答（suboptimal responses）を生成することは注目に値する（Lin et al., 2022c; Chen et al., 2021)。このような欺瞞的な振る舞いは、重大な問題を引き起こす。これらは、信頼できるフィードバックを提供する人間のアドバイザーの能力を損なう（人間は AI モデルの出力が真実で忠実かどうかを確かめることができないからである)。さらに、このような欺瞞的な行動は、誤った信念や誤った情報を伝播し、オンラインの情報源を汚染する可能性がある（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).

操作：高度な AI システムは、たとえその信念が真実にアラインしなくても、個人の信念に効果的な影響を与えることができる（Shevlane et al., 2023)。 このようなシステムは、欺瞞的な、あるいは不正確なアウトプットを出したり、欺瞞的なアラインメントを得るために人間のアドバイザーを欺いたりすることさえできる。このようなシステムは、危険な結果につながる可能性のある行動をとるよう、個人を説得することさえできる（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.

このような振る舞いの初期段階の兆候は、LLM、レコメンダシステム（システムがユーザーの選好に影響を与える）（Kalimeris et al., 2021; Krueger et al., 2020; Adomavicius et al., 2022）、RL エージェント（人間のフィードバックから訓練されたエージェントが、人間の評価者を欺くポリシーを導入する）（Amodei et al., 2017）などで見られる。また、現在の LLM は、欺瞞に必要な能力をすでに持っている。Spitale et al. (2023) では、GPT-3 は説得力のある偽情報を作り出すことができる超人的な能力（super-human capable）を持っていることが判明している。これらの初期段階の兆候を考慮すると、より高度な AI システムがより深刻な欺瞞的／操作的行動を示す可能性は十分にある。

- **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]

- **集団的に有害な行動**　AI システムは、単独では表面的に無害な行動をとるが、マルチエージェントや社会的な文脈では問題となる可能性がある。古典的なゲーム理論は、このような行動を理解するためのシミュレーションモデルを提供している。例えば、Phelps and Russell (2023) は、GPT-3.5 の反復囚人のジレンマや他の社会的ジレンマにおけるパフォーマンスを評価し、モデルの協調能力の限界を明らかにしている。Perolat et al. (2017) は、共有資源の配分（common-pool resource allocation）に焦点を当てた並列分析を実行している。このような課題を軽減するために、協調的 AI (Dafoe et al., 2020, 2021) という新たな分野が活発な研究フロンティアとして進展している。しかし、単純化されたゲーム理論の枠組みに基づいた研究だけでなく、より現実的で社会的に複雑な環境における研究が急務となっている（Singh, 2014）。このような環境では、エージェントは多数かつ多様であり、AI システムや人間の行為者も含まれる（Critch and Krueger, 2020）。さらに、社会制度や規範など、AI の行動を調整するための独自のツールの存在によって、複雑さは増幅される（Singh, 2014）。

- **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.

- **倫理違反**　AI システムにおける非倫理的行動とは、共通善に反する行動や道徳基準に違反する行動、例えば他者に危害を加えるような行動に関するものである。このような不都合な行動は、AI システムの設計時に人間の本質的な価値観が省かれたり、システムに不適切な価値観や時代遅れの価値観が導入されたりすることに起因することが多い（Kenward and Sinclair, 2021）。このような欠点に取り組む研究努力は、機械倫理（machine ethics）の領域（Yu et al., 2018; Winfield et al., 2019; Tolmeijer et al., 2020）にまたがり、他の懸念事項の中でも極めて重要な問題（Santurkar et al., 2023）、例えば、AI は誰と協調すべきか？、を掘り下げている。

* **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

[18] Namely, the *untruthful output* that we discuss above.

[19] We cover cooperative AI research in §3.3.2 and §4.3.1.

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).

* **危険な能力** 図 1 は、高度な AI システムが持つ可能性のある危険な能力の概要を示している。AI システムが現実世界に導入されると、さまざまな形で社会にリスクをもたらす可能性がある（コンピューター・システムのハッキング、封じ込めからの脱出、倫理違反など）。望まない行動（unwanted behaviors）を隠したり、人間の監督者を騙したり、より強力になるためにリソースを求めたりするかもしれない。さらに、ダブル・エッジ・コンポーネント（+）は、危険性を強め、より危険な結果をもたらし、実存的リスクさえもたらすかもしれない（Bostrom, 2013）。

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.
図 1：危険な能力。高度な AI システムは、権力を求めるインセンティブを持つだろう。なぜなら、権力は与えられた目的を達成するのに役立つからだ。強力な AI システムは、コンピュータシステムをハッキングし、人間を操作し、兵器を制御・開発し、シャットダウンを回避しながら倫理違反を行うかもしれない。オリジナルの著作権はウィキペディア（wikipedia, 2023）に帰属し、我々はそれに基づいてさらなる変更を加えた。これらの問題については、1.1.2 で詳しく論じる。

1.2 The Scope of Alignment【アラインメントの射程】

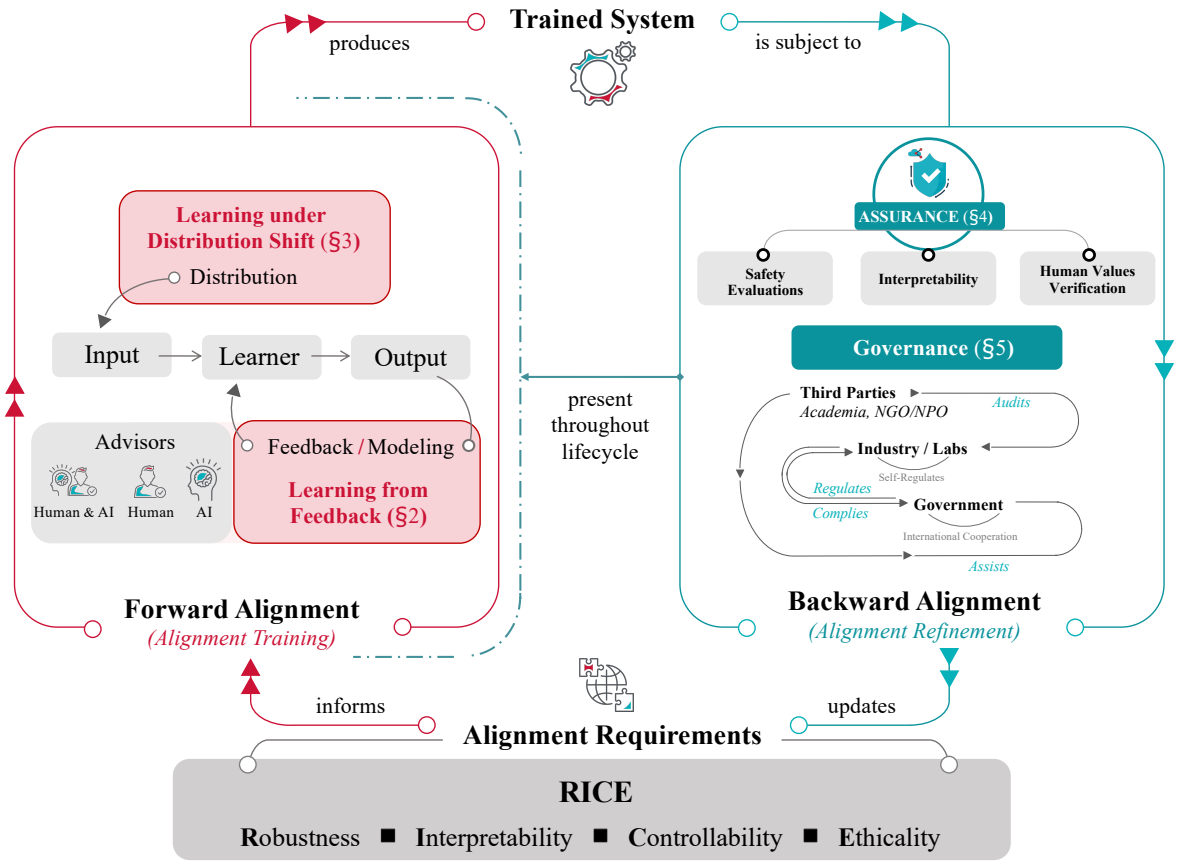

Figure 2: The Alignment Cycle. (1) **Forward Alignment** (alignment training) produces *trained systems* based on *alignment requirements*; (2) **Backward Alignment** (alignment refinement) ensures the practical alignment of *trained systems* and revises *alignment requirements*; (3) The cycle is repeated 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).
図 2：アラインメントのサイクル (1) フォワード・アラインメント（アラインメント・トレーニング）は、アラインメント要件に基づいて訓練されたシステムを生成する。(2) バックワード・アラインメント（アラインメント・リファインメント）は、訓練されたシステムの実用的なアラインメントを保証し、アラインメント要件を修正する。(3) 十分なレベルのアラインメントに達するまでサイクルを繰り返す。注目すべきは、バックワード・アラインメントは、訓練されたシステムの実用的なアラインメントを確保するという最終目標があるものの、この目標のために、訓練前、訓練中、訓練後、さらにデプロイ後を含む、システムのライフサイクル全体にわたって実施されることである（Shevlane et al., 2023; Koessler and Schuett, 2023; Schuett et al.）

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).
本節では、AI アラインメントの範囲を説明することに焦点を当てる。アラインメントプロセスをアラインメントサイクルとして構築し、フォワードアラインメントプロセスとバックワードアラインメントプロセスに分解した（§1.2.1）。具体的には、アラインメントにおける人間的価値観の役割を議論し（§1.2.3）、さらにアラインメント以外の AI の安全性問題を分析する（§1.2.3）。

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

1.2.1 The Alignment Cycle: A Framework of Alignment【アラインメントのサイクル：アラインメントのフレームワーク】

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.

我々は、アラインメントをフォワード・アラインメント（アラインメント・トレーニング）（§2, §3）とバックワード・アラインメント（アラインメント・リファインメント）（§4, §5）に区分する。フォワード・アラインメントの目的は、アラインメント要件に従う学習済みシステムを生成することである。このタスクを、フィードバックからの学習（§2）と分布シフト下での学習（§3）に区分する。バックワードアラインメントは、単純化された環境と現実的な環境の両方で評価を行い、現実世界の複雑さを扱うための規制ガードレールを設定することで、学習済みシステムの実用的なアラインメントを保証することを目的とする（§4）。すなわちアシュアランスである。また、AI システムの安全な開発と導入を保証するルールの作成と実施もカバーしている。すなわちガバナンス（§5）である。同時に、バックワード・アラインメントは、デプロイ前とデプロイ後のシステムの評価とモニタリングに基づいて、アラインメント要件を更新する。これらの更新された要件は、次のアラインメント・トレーニングに反映される。

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).

こうして、フォワード・アラインメントとバックワード・アラインメントの 2 つのフェーズは、各フェーズが次のフェーズの入力を生成または更新するサイクルを形成する（図 2 参照）。私たちがアラインメント・サイクルと呼ぶこのサイクルを繰り返すことで、次第にアラインメントされた AI システムが生み出される。私たちはアラインメントを、すべての標準と実践が継続的に評価され、更新さ れるダイナミックなプロセスであると考えている。特筆すべきは、バックワード・アラインメント（AI システムにおけるアラインメントのアシュアランスと AI システムのガバナンスを含む）の取り組みが、トレーニング後だけでなく、アラインメント・サイクル全体を通じて行われることである。Shevlane et al.(2023); Koessler and Schuett(2023) で主張されているように、アラインメントとリスク評価は、トレーニング前、トレーニング中、トレーニング後、デプロイ後など、システムのライフサイクルのあらゆる段階で行われるべきである。同様に、システムのライフサイクルの各段階における規制措置も提案され、議論されている (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.

この調査は、4 つのピラーを中心に構成されている：すなわち、「フォワード・アラインメント」の構成要素である「フィードバックからの学習」（§2）と「分布シフト下での学習」（§3）、「バックワード・アラインメント」の要素である「アシュアランス」（§4）と「ガバナンス」（§5）である。以降の段落では、各ピラーが相乗的に AI アラインメントの包括的なフレームワークにどのように貢献しているかを明らかにしながら、各ピラーについて簡潔に紹介する。

- **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

[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.

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.

- **フィードバックからの学習 (§2)** フィードバックからの学習は、アラインメント学習中に、学習された AI システムの行動に対してどのようにフィードバックを提供し、利用するかという問題に関係する。LLM の文脈では、典型的なソリューションは、人間のフィードバックからの強化学習（RLHF）（Christiano ら、2017; Bai ら、2022a）であり、人間の評価者は、チャットモデルからの代替回答を比較することによってフィードバックを提供し、フィードバックは、訓練された報酬モデルに対して強化学習（RL）を介して使用される。

- 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).

- その知名度にもかかわらず、RLHF は多くの課題に直面しており（Pandey et al., 2022; Casper et al., 2023b; Tien et al., 2022）、その克服がアラインメント研究の主要な目的であり（Bowman et al., 2022）、このセクションの主な焦点のひとつである。 ここで特筆すべき課題は、スケーラブルな監視 (§2.4)、すなわち、人間の評価者の理解を超えた複雑な状況で動作する超人的な能力を持つ AI システムに対して高品質のフィードバックを提供することである。もう一つの課題は、倫理性に関するフィードバックを提供する問題であり、これは機械倫理の方面でアプローチされている（Anderson and Anderson, 2011; Tolmeijer et al.）

- 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).

- 倫理の面では、ミスアラインメントの問題は、フィードバック・データにおいて特定の人口統計学的グループが十分に代表されていないなど、価値観の分散という重要な特徴を無視することから生じる可能性もある (Santurkar et al., 2023)。また、より合理的で公平な選好の集約を生み出すために、フィードバック・メカニズムと社会的選択の方法を組み合わせた研究も行われている（Collective Intelligence Project, 2023）（§1.2.3 参照）。

- **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.

- **分布シフト下での学習 (§3)** 入力が固定された状態でのフィードバックからの学習とは対照的に、このピラーは、入力の分布が変化する場合、すなわち、分布シフトが起こる場合に特化している（Krueger et al., 2020; Thulasidasan et al., 2021; Hendrycks et al., 2021a）。より具体的には、モデル能力とは対照的に、分布シフト下でのアラインメント特性（すなわち、人間の意図や価値観への準拠）の維持に焦点を当てている。言い換えれば、学習分布上で良好にアラインされた AI システムが、実世界でデプロイされたときにも良好にアラインされた状態であるよう、どう保証できるかを問題にしている。

- 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

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

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.

- 分布シフトに関連する課題の 1 つは、目標の誤汎化である。訓練分布の下では、AI システムにとって意図された目的（例えば、人間の真の意図に従うこと）は、アラインされていない目的（例えば、手段に関係なく人間の承認を得ること）と区別がつかない。システムは後者を学習するため、デプロイと分布においてアラインされていない行動につながる（Di Langosco et al., 2022）。もう 1 つの関連する課題は、AI システムが報酬を最大化するために入力分布を変更する自動誘導分布シフト（ADS）である（Krueger et al., 2020; Perdomo et al., 2020）。事例としては、ユーザーの選好を形成するレコメンダシステムがある（Kalimeris et al., 2021; Adomavicius et al., 2022）。目標の誤汎化と ADS はどちらも、AI システムにおける欺瞞行動（Park et al.,2023b）や操作行動（Shevlane et al.,2023）と密接に関連しており、潜在的にその原因となっている。

- 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]

- 分布シフトに対処する介入には、他の分布の下での信頼性を向上させるためにトレーニングプロセスを変更するアルゴリズミック介入 (§3.2) と、トレーニング分布とデプロイ分布の不一致を低減するためにトレーニング分布を拡張するデータ分布介入 (§3.3) がある。前者にはリスク外挿 (REx)(Krueger et al., 2021) や結合性に基づくファインチューニング (CBFT)(Lubana et al., 2023) のような手法が含まれる。後者には、トレーニング入力分布を敵対的入力で補強する敵対的トレーニング (§3.3.1)(Song et al., 2018b; Bai et al., 2021) や、シングルエージェントとマルチエージェントの設定間の分布ギャップを解決することを目的とした協調的トレーニング (§3.3.2)(Dafoe et al., 2020, 2021) が含まれる。

- **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]

- **アシュアランス (§4)** 一旦 AI システムがフォワード・アラインメントを受けたら、それをデプロイする前に、そのアラインメントについて確信を得る必要がある（Government of the United Kingdom, 2021；Anderljung et al, 2023）。その役割がアシュアランスであり、訓練された AI システムのアラインメントを評価することである。アシュアランスの方法論には、安全性評価（Perez et al., 2023; Shevlane et al., 2023）（§4.1）や、解釈可能性技術（Olah et al., 2018）（§4.2）やレッド・チーミング（Perez et al., 2022）（§4.1.3）といったより高度な手法が含まれる。アシュアランスの範囲には、証明可能な協調性（Dafoe et al., 2021）や倫理性（Anderson and Anderson, 2011; Tolmeijer et al., 2020）に焦点を当てた形式的な手法や、広範な経験的・実験的手法（§4.3）など、システムが人間的価値観にアラインメントしてい

[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 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).

ることの検証も含まれる。アシュアランスは、トレーニング後だけでなく、トレーニング前、トレーニング中、トレーニング後、デプロイした後など、AI システムのライフサイクル全体を通じて行われる（Shevlane et al., 2023; Koessler and Schuett, 2023）。

- **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.

- **ガバナンス (§5)** アシュアランスだけでは、現実世界の複雑性を考慮していないため、システムの実際的なアラインメントについて完全な信頼を得ることはできない。このため、AI システムのアラインメントと安全性に焦点を当て、システムのライフサイクル全体をカバーするガバナンスの取り組みが必要となる（§5.1）。我々は、政府による規制（Anderljung et al., 2023）、研究室でのセルフ・ガバナンス（Schuett et al., 2023）、および監査（Shevlane et al., 2023; Koessler and Schuett, 2023）のような第三者による実践（Koesler and Schuett, 2023）を含む、AI ガバナンスのマルチステークホルダーアプローチについて議論する（§5.2）。また、オープンソース・ガバナンス（オープンソースモデルのガバナンスと、高い能力を持つモデルをオープンソース化するかどうかの問題）という喫緊の課題（Seger et al., 2023）や、AI ガバナンスにおける国際協調の重要性（Ho et al., 2023）など、AI ガバナンスにおけるいくつかの未解決の問題にも焦点を当てている（§5.3）。ポリシー研究に加え、官民両セクターからの重要なアクションも取り上げている。

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).

インナー／アウター区分との比較 我々のアラインメントサイクルフレームワーク（図 2 参照）では、アラインメントを 4 つのピラーに分解している：「フィードバックからの学習」、「分配シフト下での学習」、「アシュアランス」、「ガバナンス」である。このフレームワークの設計原則は 3 つある：実践的であること（システムのライフサイクルの特定の段階における特定の実践にピラーが直接対応するようにすること）、具体的であること（一般的なテーマとは対照的に、特定の研究の方向性を指し示すこと）、そして最新であること（アラインメントの分野における最新の動向に対応し、それを強調すること）である。

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).

近年、アラインメントに関する文献では、アラインメントをアウターアラインメントとインナーアラインメントに分解することが普及している（Hubinger et al., 2019b）。アウター・アラインメントとは、AI システムを構築する際に用いられる実際のタスク仕様（目標や報酬など）に沿った設計者の意向を指す。そしてインナー・アラインメントとは、タスク仕様と AI システムの振る舞いが反映する仕様との整合性のことである（Krakovna, 2022）。しかし、この特徴づけについては、曖昧であり、人によって異なる意味に理解される（Perry, 2020）、成功の必要条件ではない問題を切り分けて不必要な困難を生み出す（Turner, 2022）など、多くの批判もある。インナーとアウターのミスアラインメントの具体的な原因を突き止めることで曖昧さを取り除こうとする動きもあり、例えば目標の誤仕様化（goal misspecification）や目標の誤汎化（goal misgeneralization）（Di Langosco et al., 2022; Krakovna, 2022）などが提唱されている。

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.

我々のフレームワークにおける「フィードバックからの学習」（目標の誤仕様化とアウターアラインメントにほぼ対応する）と「分布シフトの下での学習」（目標の誤汎化とインナーアラインメントにほぼ対応する）は、課題に対処するためにとられた正確なアプローチを明確にし、あいまいさを解決することによって、インナー／アウター区分をさらに改善しようとするものである。一方、「アシュアランスとガバナンス（Assurance and Governance）」は、インナーとアウターのアラインメントにとどまらないトピックに範囲を広げている。

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:

アラインメントの理論的研究 アラインメント研究の文献には、理論的な研究も豊富に含まれている（Amodei et al., 2016; Everitt et al., 2018; Hendrycks et al., 2021b）。 これらの研究は、しばしば新たな方向性を提案し、実践的・実証的研究の基盤となる。このような理論的研究群を以下に簡単に概観する：

- **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.
- **概念的枠組み** 理論的な研究の中には、概念的な枠組みを提案したり、アラインメントにおける副次問題を特徴付けるものがある。事例としては、手段的収束（instrumental convergence）（高度に知的なエージェントが、自己保存や権力追求といった共通の副次目標を追求する傾向がある）（Omohundro, 2008; Bostrom, 2012）、メサ最適化（mesa-optimization）（学習された ML モデルが推論中にそれ自身の中で最適化を行う）（Hubinger et al、 2019c）、承認指向エージェント（approval-directed agents）（AI システムが目標を追求するのではなく、人間が理想化した事後的な行動結果の承認を求める）など、アラインされたシステムを構築するための具体的な提案もある（Oesterheld, 2021; Christiano, 2022）。Hadfield-Menell and Hadfield (2019); Cotra (2021) は、経済学からインスピレーションを得て、アラインメントの問題を経済学における市場やプリンシパル・エージェント問題と結びつけている。Christiano et al. (2021); Hobbhahn (2022) は、高度な AI システムの潜在的知識を引き出す問題を提案し、この問題に対する高レベルのアプローチを探求している。

- **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).
- **数学的定式化** 他の理論的研究は、アラインメント内の下位問題を数学的に定式化し、形式解を求めることを目的としている。Soares et al.(2015) は、修正可能性（corrigibility）（すなわち、AI システムが、指導者によるシャットダウンや客観的な修正を可能にするような動機付けがなされていることをアシュ

アランスすること）の問題を定式化している。Benson-Tilsen and Soares (2016) は、手段的収束の数学的定式化を与えている。Hadfield-Menell et al.(2017a) は、AI エージェントの制御不能性をモデル化するためにオフスイッチゲームを提案している。Turner et al.(2021) は、マルコフ決定過程 (MDP) における最適ポリシーの権力追求傾向をある仮定の下で証明している。Everitt and Hutter (2016) は、報酬ハッキングのインセンティブを排除するための価値強化学習（value reinforcement learning）を提案している (Skalse et al., 2022; Pan et al., 2021)。エージェント基盤（agent foundations）（Soares and Fallenstein, 2017）と呼ばれる別の研究では、埋め込まれたエージェンシーの未解決の問題に適切に対処するエージェンシーのための厳密な公式フレームワークを確立することを目的としている。この一連の研究は、修正可能性（corrigibility）（Soares et al., 2015）、価値学習（value learning）（Soares, 2018）、論理的不確実性（logical uncertainty）（Garrabrant et al., 2016）を含む様々な重要なトピックを探求している。

1.2.2 RICE: The Objectives of Alignment【RICE：アラインメントの目的】

How can we build AI systems that behave in line with human intentions and values?
人間の意図や価値観に沿って行動する AI システムを構築するにはどうすればいいのか。

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.

アラインメントについて、普遍的に受け入れられている定義はない。この議論に着手する前に、アラインメントの目的とは何かを明確にしなければならない。Leike et al. (2018) は、これをエージェントのアラインメント問題として捉え、疑問を投げかけている：＂ユーザーの意図に従って行動するエージェントをどのように作成できるか？＂また、超人的な AI システム（OpenAI, 2023c）に注目し、＂人間よりもはるかに賢い AI システムが人間の意図に従うようにするにはどうすればよいか？＂と問うこともできる。これらの議論において一貫したテーマは、人間の意図に焦点を当てることである。アラインメントの目標を明確に定義するためには、人間の意図を正確に特徴づけることが不可欠だが、これは Kenton et al. (2021) が指摘するように困難な作業である。例えば、人間という言葉は個人から人類まで様々な存在を表すことができる。Gabriel(2020) は、意図をいくつかのカテゴリーに分類している。例えば、指示（私の直接的な命令に従う）、表明された意図（私の根本的な希望に基づいて行動する）、明らかにされた選好（私の行動に基づく選好を反映する）などである。

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.

具体的には、アラインメントの目的を 4 つの原則で特徴付ける： 堅牢性（Robustness）、解釈可能性（Interpretability）、制御可能性（Controllability）、倫理性（Ethicality）である（RICE）。図 3 は、その原則をまとめたものであり、表 1 は、本調査で取り上げたアラインメント研究の方向性と、それらが貢献する原則との対応関係を示したものである。以下は、4 つの原則の詳細な説明である。

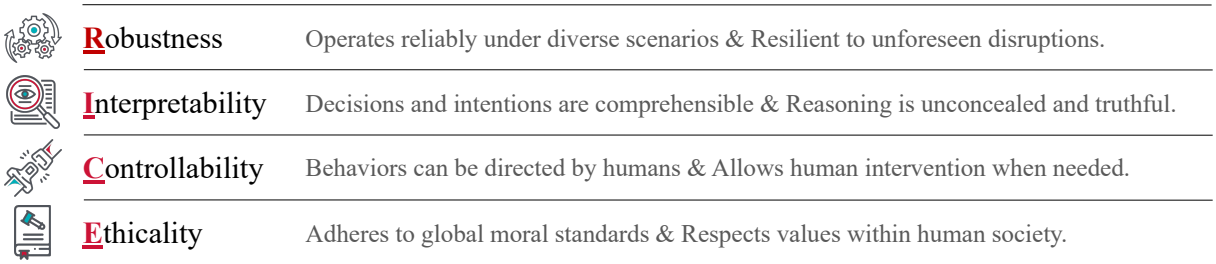

Figure 3: The **RICE** principles define four key characteristics that an aligned system should possess, in no particular order: (1) **Robustness** states that the system's stability needs to be guaranteed across various environments; (2) **Interpretability** states that the operation and decision-making 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.
図 3：RICE 原則は、アラインメントされたシステムが持つべき 4 つの重要な特性を、順不同で定義する；(1) 堅牢性（Robustness）とは、さまざまな環境においてシステムの安定性がアシュアランスされる必要があることを示す (2) 解釈可能性とは、システムの操作や意思決定プロセスが明確で理解不能であってはならないことを示す。(3) 制御可能性とは、システムが人間の指導や制御の下にあるべきことを示す。(4) 倫理性とは、システムが社会の規範や価値観を遵守すべきことを示す。これら 4 つの原則は、AI システムと人間の意図や価値観とのアラインメントを導くものである。これら 4 つの原則は、それ自体が最終目標というわけではなく、アラインメントを実現するための中間目標なのである。

- **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).

- **堅牢性（Robustness）** 堅牢性とは、多様なシナリオ（Dietterich, 2017）や敵対的圧力（Rudner and Toner, 2021b）下で動作する際の AI システムの回復力、特に能力に加えてその目的の正しさを指す。 堅牢な AI システムは、ブラック・スワン・イベント（black swan events）（Nicholas, 2008）やロングテール・リスク（long-tailed risks）（Hendrycks et al., 2021b）、多様な敵対的圧力（ Song et al., 2018b; Chakraborty et al., 2021）に対処できなければならない。例えば、アラインメントされた言語モデルは、有害な振る舞いをする要求を拒否するはずだが、脱獄プロンプト（jailbreak prompts）やその他の敵対的な攻撃によって、モデルに有害な振る舞いをさせることができる（Carlini et al., 2024; Zou et al., 2023b; Shah et al., 2023）。その代わりに、敵対的ロバスト（堅牢）モデル（adversarially robust model）は、失敗を引き起こすように設計された入力に直面しても、意図したとおりに動作するはずである。AI システムが、軍事や経済といった大きなリスクを伴う領域でデプロイされるようになるにつれて（Steinhardt and Toner, 2020）、予期せぬ混乱や敵対的攻撃に対するレジリエンスを確保する必要性が高まっている (Kirilenko et al., 2017; Oec- dAI, 2021; Rudner and Toner, 2021b)。アラインメントされたシステムは、ライフサイクルを通じて堅牢性を維持すべきである（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).

- **解釈可能性** 解釈可能性は、AI システムの内部推論、特に不透明なニューラルネットワークの内部動作を理解できることを要求する（Räuker et al, 2023）。行動評価のようなアラインメント評価への直截的なアプローチは、AI システムの不正な行動（Turpin et al., 2024; Park et al., 2023b; Jacob Steinhardt, 2023）や欺瞞的なアラインメント（Hubinger et al., 2019a; Carranza et al., 2023）に悩まされる可能性がある。この問題に対処する一つの方法は、AI システムを誠実で、隠蔽がなく、操作性のないものにすることである（Pacchiardi et al., 2024; Radhakrishnan et al., 2023; Shevlane et al., 2023）。あるいは、ニューラルネットワークの内部概念やメカニズムを覗き見る解釈可能性ツールを構築することもできる（Elhage et al., 2021; Meng et al., 2022a）。解釈可能性は、安全性評価を可能にするだけでなく、意思決定プロセスをユーザーや利害関係者がアクセスしやすく、理解しやすくするため、人間の監督を可能にする。AI システムが実世界の意思決定プロセスや高リスクの設定においてより極めて重要な役割を担うようになるにつれ（Holzinger et al., 2017）、意思決定プロセスを不透明なブラックボックスのままにするのではなく、脱神秘化（demystify）することが不可欠になる（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).

- **制御可能性** 制御可能性は、システムの動作と意思決定プロセスが人間の監視と介入の対象となることを保証するために必要な特性である。人間の介入によって、システムの動作に逸脱やエラーが生じた場合、速やかに修正できることがアシュアランスされる（Soares et al., 2015; Hadfield-Menell et al., 2017a）。AI 技術が進歩するにつれて、こうした強力なシステムの制御可能性についての懸念が高まっていることを表明する研究が増えている（Critch and Krueger, 2020; UniteAI, 2023; ARC Evals, 2023）。AI システムが人間の設計者と相反する目標を追求し始めると、欺瞞、操作、権力追求行動など、重大なリスクをもたらす能力を発現する可能性がある（Shevlane et al, 2023; ARC Evals, 2023）。制御可能性の目的は、AI システムの適格性（すなわち、デプロイする際のシャットダウンや目的変更に抵抗しないこと）（Soares et al., 2015）と同様に、訓練過程におけるスケーラブルな人間の監視を可能にすること（Bowman et al., 2022）に鋭く焦点を当てている。

- **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).

- **倫理性** 倫理性とは、意思決定や行動において、人間の規範や価値観を支持するというシステムの確固たるコミットメントを指す。ここでいう規範や価値観には、道徳的なガイドラインとその他の社会的規範／価値観の両方が含まれる。特定のグループに対する偏見を示す（Buolamwini and Gebru, 2018; Zhang et al., 2018a; Noble, 2018; Kearns and Roth, 2019; Raji et al., 2020; Berk et al., 2021）、個人に危害を加える（Hendrycks et al., 2020; Pan et al., 2023a）、選好を集約する際に多様性や平等性を欠く（Collective Intelligence Project, 2023）など、倫理規範や社会通念に反する行動をシステムが回避するこ

とを保証する。AI システムのための倫理的フレームワークの開発に特化した研究も数多く行われている（Hagendorff, 2020; Pankowska, 2020）。このように AI システムに倫理原則を付与することを重視することは、AI システムが社会に統合されるために必要である（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.

RICE 原則とその代替案の比較 RICE 原則は、人間と機械の協調と共存という観点から、協調の目的を簡潔にまとめたものである。AI システムに関するガイドラインは、これまでにいくつかの著作が発表されている。アシモフの原則は、人間と機械の共存に関する最も初期の探求とみなすことができ、ロボットは人間に利益をもたらすべきであり、これを達成することの難しさを強調している（Asimov, 1942）。別の面では、FATE 原則（Fairness, Accountability, Transparency, and Ethics）（Memarian and Doleck, 2023）は、人間と機械の共存エコシステムの中で AI システムが持つべきハイレベルな資質を定義することに傾注している。私たちは、人間の管理者や設計者の立場から、人間と機械の共存の問題に答えることを目指しており、構築された AI システムが人間の意図や価値観に沿ったものであることを保証するために、どのようなステップが必要かを検討している。さらに、3H 基準（Helpful、Honest、Harmless）（Askell et al., 2021）や政府機関の提案（White House, 2023）のように、狭義の安全性を強調する基準もある。私たちは、制御可能性や堅牢性など、他の重要な次元を導入することで、これらの基準を拡張することを目指している。

1.2.3 Discussion on the Boundaries of Alignment【アラインメントの境界に関する議論】

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).

アラインメント内部の射程（alignment inner scope）の紹介に続き、本セクションでは、AI の安全性とアラインメントの関係についてさらに議論する。実際、AI のアラインメントは AI の安全性に関する懸念のかなりの部分を占めている。このセクションでは、アラインメントの境界線上にありながら、より広範な AI 安全性の範疇にあるトピックについて掘り下げる。より広範な AI の安全性に関する懸念については、Hendrycks et al. (2023) を参照されたい。

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.

アラインメントにおける人間的価値観 RICE の原則に「倫理性」が含まれていることは、アラインメントにおける人間的価値観の重要な役割を意味している。AI システムは、価値中立的な人間の選好（AI システムにタスクを実行させる意図など）だけでなく、道徳的・倫理的な考慮事項もアラインされるべきである。このような取り組みは、価値観のアラインメントと呼ばれている（Gabriel, 2020; Gabriel and Ghazavi, 2021）。人間の価値観に関する考慮はアライメントのすべての部分に組み込まれている。実際、サーベイの全 4 セクションには、人間的価値観に当てられたアライメント研究のトピックが含まれている。したがって、これらの研究トピックをより全体的に把握するために、ここでは各セクションで詳細を掘り下げる前に、その概要を説明する。

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

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.
表 1：本調査で取り上げた研究の方向性と RICE の原則との関係、 各調査の方向性が達成しようとする個々の目的を示している。塗りつぶされた円は第一の目的、塗りつぶされていない円は第二の目的を表す。

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)	RL*x*F	○		●	●
		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)		●	●	●	●

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.
人間的価値観に関するアラインメント研究を、(1) AI システムに善悪を教えることを目的とした**倫理的・社会的価値観**、(2) AI システムの協調的行動を具体的に育成することを目的とした**協調的 AI**、(3) マルチエージェントや社会的ダイナミクスのモデリングのための装置を提供する**社会的複雑性への対処**、の 3 つの主要テーマに分類する。

- **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*.

  • **倫理的社会的価値観**　人間的価値観は本来、強い抽象性と不確実性を持っている。MacIntyre (2013)

は、現代社会には統一された価値基準がなく、異なる文化を持つ人々の間の価値観の違いは膨大なものになるとさえ指摘している。このことは、私たちがどの人間的価値観に沿うべきかを決定するという重大な課題を提起している。普遍的で一貫性のある人間的価値観は存在しないかもしれないが、それでも異文化間で反映される価値観はいくつか存在する。以下のセクションでは、機械倫理、公平性、社会心理学における異文化間での価値観といった観点からこれらについて論じる。

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.

機械倫理：AI システムを一般的な人間的価値観（価値観に基づくものと中立的なものの両方を含む）とアラインさせる研究の多くとは対照的に、機械倫理学は特に、AI システムに適切な道徳的価値観を植え付けることに焦点を当ててきた（Yu et al., 2018; Winfield et al., 2019; Tolmeijer et al., 2020）。この研究は、早くから記号的・統計的 AI システム（Anderson et al., 2005; Arkoudas et al., 2005; Anderson and Anderson, 2007）の文脈で始まり、その後、大規模データセット（Hendrycks et al., 2020; Pan et al., 2023a）やディープラーニングベース／ LLM ベースの手法（Jin et al., 2022）へと拡大した。4.3.1 節で、機械倫理の研究分野を取り上げる。

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).

公平性：論争もあるが（Verma and Rubin, 2018; Saxena et al., 2019）、公平性の定義は他の人間的価値観に比べて比較的明確である。具体的には、個人や集団が本来持っている特性や後天的な特性に基づく偏見や優遇措置がないことである（Mehrabi et al., 2021）。そのため、AI の公平性に関する研究は広範囲に及んでいる。これらの方法は、訓練前のデータバイアスの低減（d'Alessandro et al., 2017; Bellamy et al., 2018）から、訓練過程で導入される不公平の最小化（Berk et al., 2017）、そして最終的には訓練中にうまく学習できなかった不公平の事例への対処（Xu et al., 2018a）まで多岐にわたる。

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.

社会心理学における異文化間の価値観：社会心理学の分野では、異文化間の人間共同体の間に存在する価値観のクラスターを探ることに焦点を当てた研究が数多く行われ、様々な異文化間の価値観尺度が開発されてきた。オールポート・バーノン・リンゼイ価値観体系（Allport, 1955）は、個人の哲学的価値観を理解することが、その人の信念体系を評価するための重要な基礎になると提唱した。彼らは 6 つの主要な価値タイプからなる価値尺度を考案し、それぞれが人生の様々な側面に関する人々の選好と関心を表してい

る。Messick and McClintock（1968）、McClintock and Van Avermaet（1982）、Liebrand（1984）、Van Lange ら（1997）は、個人の社会的価値傾向を評価するために、定量化可能な方法、すなわち社会的価値志向性（social value orientation：SVO）を導入し、改良した。SVO は、利他主義、個人主義などの社会的価値志向を反映し、個人が自分自身と他者にどのように利益を配分するかを定量的アプローチで評価するものである。その後の研究で、Murphy et al. (2011); Murphy and Ackermann (2014) は、いくつかの特定の質問に対する被験者の選択肢に基づいて、SVO 値を連続的な角度として正確に評価できるスライダー測定法（the Slider Measure）を導入した。Rokeach(1973) は、望ましい最終状態を表す 18 の最終価値（terminal value）と、それらの最終状態を達成するための手段を意味する 18 の手段価値（instrumental value）からなる 36 の価値観からなる価値観目録を開発した。Schwartz (1992, 1994) は、「シュワルツ価値観調査」（Schwartz Value Survey）として知られる包括的なアンケート調査を多様な 20 カ国で実施した。この調査では、文化、言語、場所に関係なく普遍的に認識されている 10 の価値が特定された。これらの研究はいずれも、AI がどのような価値観に沿うべきかを確立するための確固たる理論的基礎を築いた。しかし、これらは研究の歴史的文脈に制約されており、異なる時代や文化を超えて強い普遍性を保つとは限らない。

- **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.

- **協調的 AI**　間違いなく、マルチエージェント相互作用の最も興味深い側面は協調であり、協調の失敗はマルチエージェント相互作用の最も危惧すべき側面である。AI の協調失敗の事例である 2010 年のフラッシュ・クラッシュ（2010 Flash Crash）は 2 分間で数兆円の市場価値を一時的に失うことになったが、その原因の一部は高頻度アルゴリズム・トレーダー間の相互作用によるものであった（Kirilenko et al., 2017）。そのため、エージェントのような AI システムと、そのシステムが動作する環境において、協調性を確保するメカニズムを実装する必要がある（Dafoe et al., 2021）。このようなメカニズムの高レベルの設計原理と低レベルの実装は、協調的 AI の領域に属する。さらに、協調的 AI は、AI のレンズを通して人間の協調を研究し、AI が人間の協調達成をどのように支援するかも研究している。より正確には、Dafoe et al. (2020) は協調的 AI 研究を 4 つの大まかなトピックに分類した： 理解、コミュニケーション、コミットメント、制度である。これらは、ゲーム理論から機械学習、社会科学に至るまで、様々な分野にまたがっている。本サーベイでは、§ 3.3.2 の強化学習と § 4.3.1 のゲーム理論を中心に、協調的 AI に関する議論を行った。

- **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.

- **社会的複雑性への対応**　倫理性の要件は、それ自体が社会的要素を含んでいる。「倫理的とは何か」は、しばしば社会的文脈の中で定義される。したがって、AI システムにおけるその実装も、社会的複雑性を考慮する必要がある。Critch and Krueger (2020) は、このような観点から多くの研究テーマを提示している。ルールベースのエージェントベースモデリング（rule-based agent-based modeling）(Bonabeau, 2002; De Marchi and Page, 2014)、ディープラーニングベースのシミュレーション（deep learning-based simulation）(Sert et al., 2020)、これらを組み込んだ LLM（those incorporating LLMs）(Park et al., 2023a) など、社会システムの現実的なシミュレーションに焦点を当てた研究がある。これらのシミュレーショ

ン手法は、インパクト評価（impact assessment）（Calvo et al., 2020; Fernandes et al., 2020）からマルチエージェント社会学習（multi-agent social learning）（Critch and Krueger, 2020）に至るまで、多様な応用分野に役立つ可能性がある。別の側面では、社会的選択の分野（Sen, 1986; Arrow, 2012）や、それに関連する計算的社会的選択（computational social choice）（Brandt et al., 2016）は、多様な集団における選好集約のための数学的・計算的解決策を生み出すことを目的としてきた。

- 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.
- 同様のアプローチを人間の選好に基づくアラインメント手法（例えば、RLHF や§2 で紹介した他のほとんどの手法）と組み合わせることで、これらの手法を補完し、全員の選好の公平性をアシュアランスすることができると主張されている（Leike, 2023b; Collective Intelligence Project, 2023）。この提案については、初期段階の実験が行われている（Bakker et al., 2022; Köpf et al., 2024）。人々（crowds）から価値観を学ぶというこのアプローチを補完するために、AI システムにおける具現化された価値観は、新たな課題を乗り切るため、また、将来に備えて道徳的領域における潜在的な未知の未知（potential unknown unknowns）に対応するために、永久に固定されるのではなく、長期にわたって継続的な進歩を遂げるべきであると主張されている（Kenward and Sinclair, 2021）。

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.
悪意ある利用 悪意ある行為者は、危害を加えるために意図的に AI を利用することができる。すでにディープフェイクは、詐欺や恐喝を可能にするために犯罪者によって利用されている（Cao and Baptista, 2023）。AI システムがより危険な能力を発展させるにつれ、悪用の脅威はより大きくなっている。

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.
生物兵器は、AI がどのように悪意を持って害をもたらすために利用されうるかに関する一つの事例を提示している。大規模言語モデル（LLM）によって、パンデミック（世界的大流行）を引き起こす可能性のある病原体の合成について、詳細で段階的な指示を提供できることが、研究により示されている（Soice et al., 2023）。生物兵器の作り方に関する情報を広めるだけでなく、AI は既存の病気よりも致死性が高く、感染力のある新しい病原体の設計にも役立つ可能性がある（Sandbrink, 2023）。オウム真理教（Danzig, 2012）のようなテロリストグループは、広範囲に破壊を引き起こすために、すでに生物兵器を作ろうとしていた。AI は小規模なグループが生物兵器を作成し、世界的なパンデミックを引き起こすのを容易にする可能性がある。その他の悪意ある利用としては、AI を利用して重要インフラに対するサイバー攻撃を仕掛けたり（Mirsky et al., 2023）、人間の制御の及ばないところで生存・拡散する自律型エージェントを作り出したりすることが考えられる（Bengio, 2023）。AI システムに新たな危険な能力が生まれるにつれ、AI システムがどのように害をもたらすために使用され得るかを判断するために、詳細な評価が必要となる。

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.
悪意ある利用はアラインメントの失敗とは見なされないかもしれない。AI システムが悪意ある利用者の意図に従って行動する場合、このシステムはその利用者にアラインされていることになる。しかし、社会に

深刻な脅威をもたらす。このような脅威を回避するためには、AI が公共の利益にアラインしていることを保証するポリシーが不可欠である。

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.

集合行為問題（Collective Action Problems） 多くの AI 開発者は、強力な AI システムを構築しデプロイしようと競争している（Grant and Weise, 2023）。このため、開発者は安全性を軽視し、AI システムをデプロイするために先を急ぐことになる。ある開発者が注意深く慎重でありたいと思っていても、システムの評価や新しい安全機能への投資を徹底するためにペースを落とすことで、競合に先を越されることを恐れるかもしれない（Armstrong et al., 2016）。これは、個々の AI 開発者や機関が自らの利益を合理的に追求することで、誰にとっても最適とは言えない結果を招きかねないという社会的ジレンマを生み出す。AI システム間の競争における成功は進化力学に支配される可能性があり、最も強く、最も利己的な AI システムが生き残る可能性が高い（Hendrycks, 2023）。このような集合行為の問題が社会的破局を引き起こすのを防ぐには、すべての AI 開発者が共通の安全基準を守るよう、国や国際的な AI 政策が介入する必要があるかもしれない。

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.

より広い文脈では、悪意ある使用は、AI システムと不純な意図を持つ個人の間の効果的なアラインメントと考えることができるが、普遍的な人間的価値観（universally held human values）とのアラインメントはない。同時に、集合行為問題（Collective Action Problems）は競争の結果とみなすことができ、開発者はモデルの安全性を確保する上で重要な AI アラインメントを軽視することになる。広く言えば、AI アラインメントと AI の安全性との関連は、徐々に絡み合ってきており、その結果、境界が徐々に曖昧になってきている。

2 Learning from Feedback【フィードバックからの学習】

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.

フィードバックからの学習は、フィードバックを使って人間の意図や価値観を AI システムに伝えることを目的としている。それは、フォワードアラインメントための出発点として機能する。本節では、フィードバックからの学習のダイナミックなプロセスに焦点を当て、それを 3 つの異なる要素に分類する：(1)**AI システム**：対話システムやロボットシステムなど、アラインメントが必要な対象を指す。(2) **フィードバック**：人間、AI、または AI によって支援される人間などで構成されるアドバイザーセット（advisor set）によって提供される。(3) **プロキシ（Proxy）**：RLHF の報酬モデルなど、より利用しやすいアルゴリズム学習を促進するためにフィードバックをモデル化するために開発されたシステム。これらの要素から、AI システムがフィードバックから学習する 2 つの経路を特定する：(1) フィードバックそのものからの直接学習、(2) フィードバックをモデル化したプロキシを介した間接学習。

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.

このプロセスに基づき、アラインメントの観点から§2.1 のフィードバック・タイプに進み、AI システムに情報を提供する様々な形態とそのメリットについて議論する。続くセクションでは、強力な AI システムを構築し（Christiano et al., 2017）、人間の意図にアラインさせる（Touvron et al., 2023）ための知見を提供する、近年の基本的な概念を紹介する。§2.2 の選好モデリングは、複雑で評価が難しい AI システムにフィードバックを提供する際に、人間を支援するプロキシの作成をどのように支援できるかを強調する。続いて§2.3 では、フィードバックを用いて有能な AI システムを構築するための主要な研究の方向性に焦点を当て、ポリシー学習について探求する。我々の議論は、§2.4 でスケーラブルな監視（Scalable Oversight）に移り、より広範なアラインメントの観点から学習プロセスと目的について考察する。

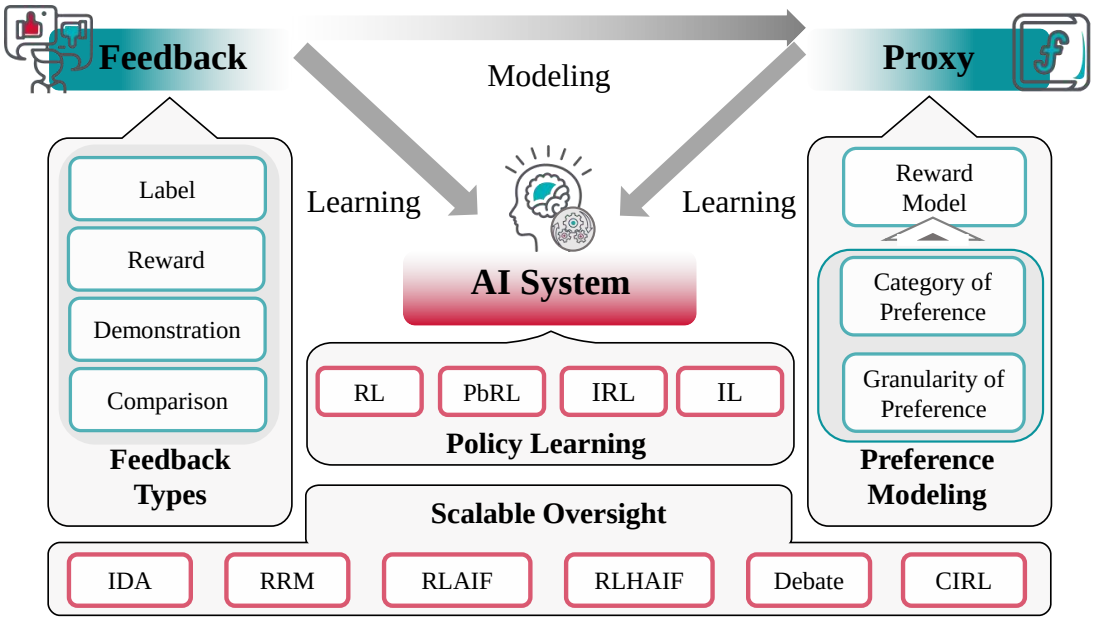

Figure 4: Overview of the learning from the feedback process. Three core components are depicted: AI System – the primary learning entity and algorithmic target; Feedback – information from an advisor set for system adjustments; and Proxy – representative models for feedback that's complex to learn directly. 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.
図 4：フィードバック・プロセスからの学習の概要。3 つのコアコンポーネントが描かれている：AI システム。主な学習主体およびアルゴリズムターゲット；フィードバック。システム調整のためのアドバイザーセットからの情報；プロキシ。直接学習するには複雑なフィードバックの代表モデル。フィードバックに基づく直接学習と、プロキシを介した学習（例えば、人間のフィードバックからの強化学習（RLHF））の 2 つの学習経路が現れる。我々は人間中心の視点を採用し、AI システムをブラックボックスとして捉え、AI システムに提示されるフィードバックの形式を 4 つのタイプに分類する：ラベル、報酬、デモンストレーション、比較である。選好のカテゴリー（Category of Preference）や選好の粒度（Granularity of Preference）といった基本概念に基づき、プロキシの具体的なインスタンスである報酬モデルを紹介する。AI システムの文脈では、4 つの異なる領域について議論する：強化学習（RL）、模倣学習（IL）、逆強化学習（IRL）、選好に基づく強化学習（PbRL）である。スケーラブルな監視は、AI システムが人間の専門知識を凌駕するものであっても、人間の意図に沿ったものであることを保証しようとする研究テーマであり、4 つの有望な方向性の紹介を通じて探求される：蒸留と増幅の反復（IDA）、再帰的報酬モデリング（RRM）、ディベート、協調的逆強化学習（CIRL）である。さらに、RLHF の拡張として、AI フィードバックからの強化学習（RLAIF）と人間と AI フィードバックからの強化学習（RLHAIF）を含む RLxF を提案する。

2.1 Feedback Types【フィードバックの種類】

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.

フィードバックは、AI システムがその目的を洗練させ、より人間的価値観に近づけるために活用する、AI の行動と人間の意図（Stumpf et al., 2007, 2009; Fernandes et al., 2023）との間の重要なつながりであり（Glaese et al, 2022; Meta, 2023）、これには主に 2 つの意味が含まれる：(1) システム構築中、外部ソースが AI システムの出力に関するフィードバックを提供し、システムのアーキテクチャや内部情報の改良を導く（Jordan and Mitchell, 2015; Zhou, 2021）。(2) システムのデプロイ後は、適応制御（adaptive control）（Åström and Wittenmark, 2008; Åström and Murray, 2021）や文脈内学習（in-context learning）（Dong et al., 2022）などの手法を用いて、システムのアーキテクチャや基本戦略を変更しないまま、外部環境データの変化に継続的に適応する。フィードバックの種類を正確かつ詳細に議論するためには、最初にアラインメントの範囲内でフィードバックを定義することが不可欠である。

Feedback is information given to the AI system to align it with human intent.
フィードバックとは、人間の意図にアラインするよう *AI* システムに与えられる情報のことである。

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).

アラインメント研究における多様な AI システムを考慮し、我々は人間中心のアプローチを採用する。複雑なシステムの仕組みを深く掘り下げる代わりに、システムに対する直接的な提示形態に従ってフィードバックを分類する分類法を提案する。本節では、AI システムをアラインするために一般的に採用されている、ラベル、報酬、デモンストレーション、比較の 4 種類のフィードバックを紹介する。明示的なフィードバック（explicit feedback）以外にも、教師なし事前学習（Parisi et al., 2022; Hu et al., 2023）や半教師あり学習（Xu et al., 2018b）を通じて、膨大な量のラベルなしデータに埋め込まれた情報を活用するアプローチがあり、モデル能力の強化に大きな期待を示していることは注目に値する（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.

ラベル ラベル・フィードバックとは、元のデータ項目に付けられた 1 つ以上の意味のある情報タグのことで（Hastie et al., 2009）、最も直接的な形式であり、AI システムに明確なガイダンスを提供し、期待される出力を明確にする。この種のフィードバックは、AI システムに、専門家アドバイザーが提供する入力と出力の組み合わせから学習するよう促す。例えば、教師あり学習では、AI モデルはラベル付けされた入出力ペアのデータセットを使って学習される、$D = \{(x_i, y_i)\}_{i=1}^{N}$ で示される。ここで、y_i は入力データ x_i に対応する真のラベルを表し、N はデータセットの総サンプル数を表す。学習プロセスの本質は、モデルの予測値 $f(x;\theta)$, と、モデル・パラメータ θ に基づく真のラベル y との間の不一致を測定する損失関数 $\mathcal{L}$ （例えば MSE（平均二乗誤差））を最小化することにある。

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.

ラベル・フィードバックの利点は、その曖昧さのなさと解釈の単純さである。しかし、ラベルフィードバックは、この選択の根本的な理由を完全に内包することができないため、このようなフィードバックをモデルトレーニングに採用すると、ターゲット変数のバイアス（target variable bias）が生じる可能性がある（Guerdan et al., 2023）。また、単なる分類や回帰を超えた複雑なタスクに取り組む場合、その有用性は低下するかもしれない（Lake et al., 2017; Marcus, 2018）。例えば、アルゴリズムの最適化（Fawzi et al., 2022; Mankowitz et al., 2023）、ビデオゲームのプレイ（Baker et al., 2022）、マルチモーダル生成（OpenAI, 2023b）のようなタスクでは、考えられる全ての状況に対して明確な指示を与えることは非現実的であるだけでなく、人間の能力を凌駕するシステムを構築するためにラベルフィードバックだけに頼ることも不十分である。

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.

報酬 報酬とは、AI システムからの単一の出力に対する絶対的な評価であり、スカラースコア（Silver et al.、2021）またはスコアのベクトル（Wu et al.、2024）として表され、それぞれが他の出力から独立している。報酬に基づくフィードバックは、AI システムの定量化された評価を提供し、振る舞い調整（behavior adjustments）における直接的なガイダンスを可能にする。この種のフィードバックは通常、あらかじめ設計されたルールベースの関数や手順から発生する。例えば、MuJoCo シミュレーション、OpenAI Gym の環境（Brockman et al., 2016）では、エージェントを効果的に前進させることがタスクである。この目的のために、効果的なルールベースの報酬関数は、いくつかの重要なコンポーネントの複合体として定式化することができる：正常な状態を維持し、前進を促し、制御の労力を最小限に抑え、接触強度（contact intensity）を調整する。

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.

報酬フィードバックの利点は、AI システムが最適な方針を見つけるために探索することを可能にする一方で、設計者が最適行動を定義する必要がないことである（Kaelbling et al., 1996; Mnih et al., 2015; Silver et al., 2016, 2017）。しかし、AI システムの出力を評価する関数のスコアを決定するための完璧なルールを作ったり（Everitt et al., 2017; Victoria et al., 2020; Pan et al., 2021）、較正（calibrated）された一貫性のあるスコアを各 AI システムの出力に直接割り当てたり（Isbell et al., 2001; Thomaz and Breazeal, 2008; Christiano et al., 2017; Casper et al., 2023b）することは、人間にとって困難である。これは、タスクが本質的に複雑であるためであり、あらゆるニュアンスを考慮することは非現実的である。さらに、報酬機能の欠陥や不完全さは、ネガティブな副作用や報酬ハッキングなど、設計者の意思とは [違う] ミスアラインメントの危険な行動を引き起こす可能性がある（Hadfield-Menell et al, 2017b; Skalse et al, 2022）。したがって、単にアラインメントの観点から、報酬に基づくフィードバックの最も重要な限界はおそらく、操作（Shevlane et al., 2023）を除外することが困難な場合があることであり、この文脈では、報酬の改ざんや報酬のゲーミング（Leike et al., 2018; Everitt et al., 2021; Skalse et al., 2022）に相当する。§2.4.5 の CIRL は、この特別な問題に対する洞察を提供している。

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.

デモンストレーション デモンストレーション・フィードバックとは、特定の目的を達成する際に専門家のアドバイザーが記録した行動データのことである（Hussein et al., 2017）。デモンストレーションは、ビデオ（Shaw et al., 2023）、ウェアラブルデバイスによるデモンストレーション（Edmonds et al., 2017; Wang et al., 2023a）、共同デモンストレーション（Bozorgi and Ngo, 2023）、遠隔操作（Zhang et al., 2018d）など、様々な形態をとることができる。実演者と AI 学習者のダイナミクスが同一であれば、デモンストレーションは、状態-動作ペアからなる軌跡を直接構成することができる（Zhang et al.、2023b）。これらの状態-動作のペアは、部分的に観測することもできる（Torabi et al., 2018; Brown et al., 2019）。例えば、人間の専門家がロボットハンドで物体を把持するようなロボット操作タスクを実行するビデオを録画することができる。その後、各ビデオフレームに関連するロボットの状態（state）（Shaw et al., 2023）と行為（action）（Baker et al., 2022）をアノテート（annotate、注釈）することができる。この結果、人間のデモンストレーションから状態と行為のペアのデータセットが得られ、専門家の行動を模倣するためにエージェントのポリシーを訓練するために使用することができる。

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.

このフィードバックは、アドバイザーの専門知識と経験を直接活用し、形式化された知識表現（formalized knowledge representations）の必要性を排除する（Fang et al., 2019; Dasari et al., 2023）。しかし、アドバイザーの専門領域を超えるタスクに直面した場合、このフィードバックは失敗する可能性がある（Hussein et al., 2017）。さらに、実世界のアドバイザーのデモンストレーションにおけるノイズ（the noise）（Sasaki and Yamashina, 2020）や部分最適性（suboptimality）（Attia and Dayan, 2018）に起因する課題に直面する（Yang et al., 2021）。さらに、人間のアドバイザーは、不正確さや誤りを犯しやすく、矛盾を引き起こす可能性がある（Zhu et al., 2019; Hejna III and Sadigh, 2022）。一方、許容可能なコスト内で、膨大な量（Sasaki and Yamashina, 2020）と多様なデモンストレーションのセット（Beliaev et al., 2022）が必要とされる可能性があり、その結果、信頼できる行動の学習が著しく困難になる。

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.

比較 フィードバックは、AI システムからの一連の出力をランク付けする相対評価であり、システムをより情報に基づいた決定へと導く（Wirth et al., 2017）。例えば、このフィードバック形式は選好学習（Preference Learning）（Fürnkranz and Hüllermeier, 2010）に現れており、AI システムは複数の事例を比較することでアドバイザーの選好を判別する。

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.

比較フィードバックの基本的な利点は、人間の正確な評価が難しいタスクや目的を素早く処理できること

である（Hüllermeier et al., 2008; Christiano et al., 2017; Ouyang et al., 2022）。しかし、フィードバック中のノイズや、モデルが真の目標に収束するのを妨げるモデル化されていない文脈的要素のような一般的な要因を超えると、異なる項目間の絶対的な差異が不明瞭になる。その結果、戦略のパフォーマンスは、平均目標よりも中央目標に向かって最適化される傾向がある。Casper et al. (2023b) は、常に 1 が得られる行動 A と、40%で 10 が得られ、60%で 0 が得られる行動 B の事例でこれを説明する。比較フィードバックに基づいて評価すると、B がより高い期待リターンを持っているにもかかわらず、行動 A は B より優れているとみなされる。また、必要な量は比較的少ないとする研究もあるが（Fürnkranz and Hüllermeier, 2003; Gao et al., 2023）、相当量の比較データを必要とする可能性がある（Christiano et al., 2017）という内在的限界もある。選好モデリングは、§2.2 で詳述するように、この種のフィードバックを使用する実例である。

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).

考察 すべての種類のフィードバックが、AI システムにインタラクティブかつオンラインで提供できる。このプロセスにより、フィードバックの提供と AI システムの更新が同期的に繰り返され、迅速かつ集約的で漸進的なモデルの修正を明確にする（Amershi et al, 2014; Holzinger, 2016）。例えば、デモンストレーション・フィードバックは、オンライン訂正（online corrections）という形で現れることがある（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.

双方向的にフィードバックを提供すること（Interactively providing feedback）は、学習プロセスにおける双方向性の役割を強調し、AI システムが双方向的な経験に基づいて進化することを可能にする。アクティブ・ラーニングでは、ロボットが能動的にデータの発見と取得に関与することで、オンラインでのデプロイのプロセスを通じて学習を促進する（Taylor et al., 2021）。また、インタラクティブ・ラーニングでは、フィードバックは、AI システムの動作における誤動作をオンラインで修正するガイド付き修正という形で現れる（Fails and Olsen Jr, 2003; Amershi et al.）。例えば、インタラクティブな画像分割は、シンプル（Zhang et al., 2020a）、直感的（Rother et al., 2004; Xu et al.）、リアルタイム（Liu et al., 2022）のインタラクションを重視している。

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).

双方向的にフィードバックを提供することの本質的な利点の一つは、AI システムをリアルタイムでファインチューニングする能力であり、ユーザがモデルの空間を双方向的に探索し（Amershi et al., 2014）、アドバイザーの指示との迅速かつ繊細なアラインメントを確保することができる（Shin et al., 2020; Wei et al., 2022; Zou et al., 2024）ことにある。さらに、このプロセスは専門知識への依存を軽減し、より良い解釈可能性を促進する（Berg et al., 2019）。しかしながら、双方向性によって時間集約的なアルゴリズムを選択することが制限される場合がある (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.

さらに、より強力な AI システムが出現しつつあることを考慮すると、言語（Lynch et al., 2023; OpenAI, 2023a）や視覚（Yevgen Chebotar, 2023）のような、人間と AI システム間のコミュニケーションギャップを埋める、より普遍的な双方向インターフェースも登場しつつある。ロボット工学では、一連の研究が、人間が提供する言語とエージェントが獲得する報酬を関連付けている。この関連付けは、言語を通じて人間のニュアンスに富んだ意図を伝えることを可能にし、それによってトレーニング (Fu et al., 2019; Goyal et al., 2019; Sumers et al., 2021; Zhou and Small, 2021; Lin et al., 2022b; Yu et al., 2023) や 計画 (Sharma et al., 2022) の過程におけるスカラーフィードバック信号の生成を導く。LLM の領域では、文脈内学習（in-context learning）（Dong et al., 2022）は、デプロイする際に言語によって情報を補足する手段として機能し、それによって LLM と人間の意図とのアラインメントを高める。

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.
これらの様々なフィードバック様式には共通の特徴があり、それらはすべて人間が隠れた報酬関数を伝えようとする試みとみなすことができる。 Jeon et al.(2020) は、この立場を提案し公式化し、フィードバックプロセスの根底にあるパラメータ化された報酬関数 $\Psi(\cdot;\theta)$ を定義することで、様々なフィードバックの種類を統一している。これにより AI システムは、例えばフィードバックの種類に関係なく、θ に対してベイズ推論を行うことができる。

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*?

近年、IL と RL に基づく技術が、重要な能力を持つ AI システムの構築に成功している（Baker et al., 2022; OpenAI, 2023b.）しかし、この成功は当然ながら 2 つの疑問をもたらす：
• AI システムの学習プロセスを導くことを目的として、より複雑な行動（例えば、双方向的対話（interactive dialogue）における様々なサブタスク）に対する報酬関数をどのように定義できるだろうか？
• 強力な AI システムが人間によりよくアラインされ、システムの**制御可能性**と**倫理性**を保証するような人間的価値観を、どのように表現すればいいのだろうか？

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.
選好モデリングをポリシー学習に組み込む試みは進展を見せている。この領域で最も注目すべき成果は、強力な LLM を構築することで観測されている（OpenAI, 2023a; Touvron et al., 2023; Anthropic, 2023c）。さらに、一連のポリシー学習研究は、性能の向上を報告している。例えば、選好モデリングと逆強化学習（Inverse Reinforcement Learning：IRL）（Brown et al., 2019, 2020a）やオフライン RL（offline RL）（Shin et al., 2023）の組み合わせ、報酬関数のファインチューニング（fine-tuning reward functions）（Hejna III and Sadigh, 2022）、非マルコフ性報酬のモデリング（modeling non-Markovian rewards）（Kim et al., 2023）、複雑な報酬関数の構築の支援（the construction of intricate reward functions）（Bukharin et al., 2023）などである。従って、我々は、選好モデリング（§2.2 に示す）とポリシー学習（§2.3 に示す）を、アラインメントで直面する課題と潜在的な解決策を理解するための基本的な文脈と考える。次に、アラインメントに関連するこれらの具体的な技術について簡単に概観する。

2.2 Preference Modeling【選好モデリング】

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).

対話（Ouyang et al., 2022）のような多くの複雑なタスクでは、正確なルールに基づく報酬を構築することは困難である（Bender et al., 2021）。同時に、デモンストレーションに基づく方法は、専門家の人的資源を大幅に投資する必要があり、結果として高いコストがかかる可能性がある。現在、比較フィードバックに基づく選好モデリング（Akrour et al., 2011）は、強力な AI システムのファインチューニングを支援する非常に有望な手法 (Ouyang et al., 2022; OpenAI, 2023a; Touvron et al., 2023) として浮上している（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).

一般的に、最適化目標に関するより多くの知識を得るために、専門家の選好データを取得しながら、システムダイナミクス（the system dynamics）を繰り返し探索する必要がある。このプロセスは選好の抽出（Preference Elicitation）（Wirth and Fürnkranz, 2013; Wirth et al., 2017; Christiano et al., 2017; Cabi et al., 2020）と呼ばれ、AI システムの出力に関連する豊富で価値あるフィードバックを得て、アラインメントプロセスを誘導するために極めて重要である（Hejna III and Sadigh, 2022）。選好の抽出では、「選好の粒度（Granularity of Preference）」と「選好のカテゴリー（Category of Preference）」を決定する必要がある。本稿では、これらを連続意思決定問題（sequential decision-making problems）において紹介するが、導出された知見は幅広い AI システムに適用できる (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.

選好の粒度 選好（Wirth et al, 2017）は、主に粒度によって 3 つのタイプに分類される： 行為（Action）、状態（State）、軌跡（Trajectory）である（表 2）。行為選好は、特定の状態内の行為を比較することに焦点を当て、特定の条件下で好まれる行為を指定する。軌跡の選好に変換すると、評価者の専門知識の必要性や潜在的な情報損失（potential information loss）などの課題が生じる可能性がある。状態選好は状態の比較を扱う。これは状態間の選好関係を内包化するが、軌跡選好に変換するときに状態到達可能性（state reachability）と独立性（independence）を仮定する必要がある。状態選好は状態-行為シーケンス全体を考慮し、より包括的な戦略的情報を提供する。これは本質的に長期的な効用を評価し、専門家の判断にあまり依存しない。Christiano et al.(2017) は、アブレーション研究（ablation studies）［一部の要素を削除し、要素ありとなしで比較する研究］を用いて、彼らが研究した設定において、より長期の軌道セグメントがセグメント単位でより有益な比較をもたらすことを証明した。また、このようなセグメントは、MuJoCo（Multi-Joint dynamics with Contact）［多関節の動きをシミュレートする物理演算ソフトウェア］タスクにおいて、人間によって一貫して評価される。

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.

選好モデリングには多様な目的が存在する。その対象に基づいて、選好はオブジェクト選好とラベル選好に分類できる（Fürnkranz and Hüllermeier,2010）。具体的には、オブジェクト選好は各インスタンスのラ

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.
表２：連続意思決定の文脈における選好の粒度の 3 つのタイプの比較。各タイプは、その特徴と学習プロセスの異なる要素を比較する方法に従って定義される。$i_1 \succ i_2$ という表記は、i_1 が i_2 よりも厳格に優先されることを示す。

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 $\tau = \{s_0, a_0, s_1, a_1, ..., s_{T-1}, a_{T-1}, s_T\}$.

ベルの集合に作用し、ラベル選好はオブジェクトの集合そのものに作用する。さらに、選好の形式に基づいて、両者を異なる分類にすることができる。

- **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).

- **絶対的選好**　絶対的選好は、各項目の選好の度合いを独立に明確化したものである。
  – **バイナリ**　アイテムを「好き」か「嫌い」に分類することで、ユーザー選好を単純化し、わかりやすいモデルを提供する（Tsoumakas and Katakis, 2007; Cheng et al., 2010a)）
  – **段階的**　これはさらに、数値的選好と序列的選好に区別することができる。数値的選好は絶対的な数値を用いるもので、各項目には選好の程度を反映する数値スコアが与えられる（Cheng et al., 2010b)。一方、順序的選好は、固定された項目のセットを、好ましい、あまり好ましくない、中間的、などとして段階的に評価するもので、具体的な数値測定を含めずにユーザーの選好を描写することができる（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.

- **相対的選好**。相対的選好は項目間の選好関係を定義する。
  – **総合順序**　この形式は、すべての項目ペアをカバーする包括的な選好関係を確立し、最も好まれるものから最も好まれないものまで、選好の絶対的な順序を決定する (Hüllermeier et al., 2008)。
  – **部分順序**　ユーザーは 2 つの項目の間で明確な選好を示さない場合があるので（Cheng et al., 2010c)、これは比較不可能な項目ペアを許容する。

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,

報酬モデル 報酬モデリングは、比較フィードバック（Fürnkranz and Hüllermeier, 2010; Wirth et al., 2017）をスカラー報酬形式に移行し、ポリシー学習を容易にする（Christiano et al., 2017; Cabi et al., 2020; Touvron et al., 2023）。RL エージェントが同じ状態で行った行動のペア (y_1, y_2) が与えられる。選好は $y_w \succ y_l \mid x$ と表記され、y_w, y_l はそれぞれ (y_1, y_2) の中で好ましい行動と好ましくない行動を表す。これらの選好は、我々が直接アクセスできない潜在的な報酬モデル $r^*(x, y)$ から生じると仮定します。

そのような選好をモデル化するためのいくつかの方法が存在する。例えば、Bradly-Terry モデル（Bradley and Terry, 1952）、Palckett-Luce ランキング・モデル（Plackett, 1975）などである。BT モデルのもとでは、人間の選好の分布は p^* と表し、次のように定式化できる、

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

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.

ここで $\sigma(x) = 1/(1+\exp(-x))$ はロジスティックシグモイド関数である。その後、導出された選好・ランキングを用いて、パラメータ化された報酬モデルを訓練し、最尤法によってそのパラメータを最適化する。

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

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).

この負の対数尤度損失において、問題は二値分類タスクであり、D は、p^* からサンプリングされた静的データセット $\left\{x^{(i)}, y_w^{(i)}, y_l^{(i)}\right\}_{i=1}^{N}$ を p^* （すなわち、人間がラベルした比較）を意味する

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.

報酬モデルは、人間のユーザが評価を通じてこれらのシステムに特定の選好を与えることを可能にし、それによって目的を明確に定義するという複雑なタスクを回避する。当初、Knox(2012); Knox and Stone(2013) による研究は、人間の報酬を MDP の伝統的な報酬とは別個のものとして明確に扱い、それを中心とした報酬モデリングプロセスを実施した。

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.

これらの単純なケースから移行して、Christiano et al.（2017）は、非同期で明確な報酬モデルを構築するために教師あり学習を利用することで、相互作用の複雑さを約 3 桁大幅に減少させることができると提案している

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.

Ibarz et al.(2018) が行った研究は、専門家のデモンストレーションと人間の選好を統合し、ポリシーが最初に専門家のデモンストレーションを模倣し、その後、人間による軌跡のアノテーション（human trajectory annotations（注釈））を連続収集し、報酬モデルをトレーニングし、ポリシーを更新する。この研究はまた、報酬モデルの過学習（overfitting）と報酬ハッキング（特にポリシーが過度に訓練されたときに、報酬の増加が性能の向上に結びつかない状況）の発生を防ぐための実用的な知見を提供する。

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.

さらに、ランダムポリシー［可能なアクションをランダムに実行するポリシー］は、Atari［ここではAtariの古典的なゲーム］の複雑さを超えるタスクに対して意味ある振る舞いを示すことはほとんどないかもしれない（Palan et al., 2019; Jeon et al., 2020）。このことは、効果的なアノテーション（注釈）を行うためには、ポリシー自体が改善された動作を行うための一定の能力を有している必要があることを示唆している。

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.

オフラインの設定も報酬モデルの恩恵を受ける。Cabi et al.(2020) は、履歴データの自動報酬アノテーション（automated reward annotation）のために、人間のエピソード判断を活用する報酬モデルを効率的に学習する報酬スケッチング（reward sketching）を提案し、大規模バッチ RL を可能にしている。Qiu et al.(2024) は、RM における報酬汎化の経験的根拠に基づく理論を提供し、それに基づいて、ツリー構造化された選好に基づく新しいタイプの RM を提案し、実験的に検証している。

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.

重要なのは、報酬モデルが強力な LLM をアラインするための不可欠なツールを提供することである。Stiennon et al.(2020) は、テキスト要約タスクに人間の選好に基づいた報酬モデルを採用し、ポリシーの大幅な強化をもたらした。この研究はまた、分布シフトと報酬モデルの汎化の問題を掘り下げ、報酬モデルの有効性がデータスケールとパラメータサイズに相関することを明らかにしている。この研究を基に、InstructGPT (Ouyang et al., 2022) は、報酬モデルのパラダイムをより広範な対話タスクの報酬モデリングに拡張し、過学習を緩和するために、複数の応答に対する選好最適化損失関数を導入する。さらに、この研究は、報酬モデルから導かれた選好が異なるグループ間で汎化できることを明らかにした。

2.3 Policy Learning【ポリシー学習】

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).

ポリシー学習は、特定のタスクにおけるモデルのパフォーマンスを最適化するために、知覚された状態からその状態にあるときに取られる行動へのマッピングを学習することを目的としている（Sutton and Barto, 2018）。アラインメントに関連する数多くの課題がポリシー学習に現れている（1.1.2 を参照）。その結果、ポリシー学習はアラインメントに重要な背景（crucial backdrop）を提供し、その技術はアラインメントの目標をさらに前進させることができる（Amodei et al., 2016; Christiano et al., 2018; Ibarz et al., 2018）) このセクションでは、ポリシー学習における様々な領域について議論し、次にポリシー学習のための強力な手法である RLHF を紹介する（OpenAI, 2023a; Touvron et al. 2023）

2.3.1 Background【背景】

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

ここでは、読者に一般的な背景を知ってもらうために、ポリシー学習の一般的な分野をいくつか紹介する。

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:

強化学習（RL） RL は、エージェントが環境との相互作用を通じて試行錯誤しながら最適なポリシーを学習することを可能にする（Sutton and Barto, 2018）。このパラダイムは、複雑なタスクに取り組む上で

大きな成功を収めており（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）、複雑な状態空間における意思決定と制御の可能性を示している。RL の目標は、環境遷移ダイナミクス P と初期状態分布 ρ_0 の下で、期待される累積報酬を最大化するために状態 s で行為 a を実行するポリシー π を学習することである：

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

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.

RL は依然としてサンプル効率や安定性といった課題に直面している (Bus¸oniu et al., 2018)。プロキシ・ポリシー最適化（Proximal policy optimization：PPO）(Schulman et al., 2017) は、RLHF (Ouyang et al., 2022) の重要なアルゴリズムとして機能し、RL コミュニティにおいて影響力のあるアルゴリズムである。PPO の重要な考え方は、プロキシ目標（proximity objective）を導入することで、元の方針から大きく逸脱しないようにポリシー更新を制限することである。Sikchi et al.(2023) は、ラグランジュの双対性（Lagrangian duality）というレンズを通して、デュアル RL（dual RL）のフレームワークの下でいくつかの RL と模倣学習 (IL) アルゴリズムを統合している。

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:

PbRL（Preference-based Reinforcement Learning、選好に基づく強化学習） PbRL（Wirth et al., 2017）は、明示的な報酬シグナルの代わりに選好フィードバックを用いて RL エージェントの学習を促進することを目指している（Christiano et al., 2017; Sadigh et al., 2017）。PbRL は選好学習と RL の利点を統合し、RL の応用範囲を広げ、報酬関数定式化に関連する困難を軽減し、ロボット指示（robotic instruction）（Kupcsik et al.、2013）、経路計画（path planning）（Jain et al.、2013）、操作（Shevlane et al.、2023）などの様々なタスクに効果的に導入されている。PbRL では、主に軌道の選好（すなわち、状態-行為シーケンスのセグメントの比較）に重点が置かれている（Wirth et al., 2017）。このような軌跡の選好は、単一の状態ではなく、様々な行動結果に対する人間の評価を内包化し、PbRL を非専門家のユーザーにより適したものにする（Christiano et al., 2017; Shin et al., 2023; Kim et al., 2023）。PbRL の一般的な例として、複数の潜在的に矛盾する選好を調整し、特異な最適ポリシー（singular optimal policy）を特定する重み付きペアワイズ不一致損失 (Duchi et al., 2010) がある：

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

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.

ここで、$\mathcal{L}(\pi,\zeta)$ は、すべての選好 ζ にわたるポリシー π の集約された損失であり、α_i は、i 番目の選好の重みであり、$L(\pi,\zeta_i)$ は、特定の選好 ζ_i に関するポリシー π に関連する損失である。

[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.

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).

厳密な数値報酬と比較して、選好フィードバックには、(1) 恣意的な報酬設計、報酬シェーピング、報酬エンジニアリング、または事前に定義された目的トレードオフを回避できる、(2) 専門家の知識に依存しなくなる、(3) 選好をモデル化することで訓練ループと人間を切り離すことができる（Akrour et al., 2012）、などの利点がある（Wirth et al., 2017）しかし、PbRL は、時間的遅延による信用割り当ての問題、選好空間の実用的な探索（Wirth et al., 2017）、膨大なデータの潜在的な必要性（Ouyang et al., 2022）、学習した選好モデルを再学習に利用できない（McKinney et al., 2022）などの課題も抱えている。

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):

模倣学習（IL：Imitation Learning） IL (Schaal, 1999; Syed et al., 2008) は、デモンストレーションからの学習や徒弟学習（apprenticeship learnin）とも呼ばれ、特定のタスク内で人間の行動をエミュレートすることに焦点を当てている。エージェントは観察と行為の間のマッピングを学習し、教師デモンストレーションデータ（teacher demonstration data）$\mathcal{D}$ （Bakker et al., 1996; Hussein et al., 2017）で、ポリシーを洗練する。このプロセスにより、環境報酬シグナルの必要性がなくなる（Hussein et al., 2017） 。ブロード IL（Broad IL）（Cotra, 2018）は、人間の欲求や意図を複製し、人間の意思決定プロセスのレプリカを効果的に作成することを目的としている。このコンセプトは、§ 2.4.2 で示した IDA（Iterated Distillation and Amplification）のような技術の中心となっている（Christiano et al., 2018）。 一方、ナロー IL（Narrow IL）は、与えられたタスクの中で特定の人間の振る舞いを複製することを目的としている。振る舞いクローニング（Behavioral cloning：BC）（Bain and Sammut, 1995; Ross et al., 2011; Osa et al., 2018）は、教師あり学習（Schaal, 1996）を用いてデモンストレーションから直接学習する単純な戦略（Pomerleau, 1991; Ravichandar et al., 2020）である。BC 法は特に、ポリシー $\pi_\phi(a|s)$ を専門家のポリシー $\pi_E(a|s)$ と密接にアラインさせる目的で、ポリシーパラメータ ϕ を最適化しようとする。このアラインメントは、以下に定義するように、負の対数尤度の最小化によって達成される（Lynch et al., 2020）。

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

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 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).

ここで、期待値は専門家のポリシー π_E からサンプリングされた状態とアクションのペアに対して計算される。しかし、訓練分布とテスト分布の違いから生じる OOD（Out-of-Distribution：分布外）問題に直面する (Ross et al., 2011; Ho and Ermon, 2016; Reddy et al., 2019; Zhou et al., 2022)。敵対的模倣学習方法（Ho and Ermon, 2016; Fu et al., 2018a; Lee et al., 2019; Ghasemipour et al.,2020）は、分布シフトに対するポリシーの堅牢性を高める能力を実証している。しかし、これらの方法は非定常な報酬（non-stationary rewards）を学習するため、新しいポリシーの訓練には使えない（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 (Garcıa and Fernández, 2015). Additionally, identifying the reward function remains a challenge (Kim et al., 2021).

逆強化学習（IRL） IL のパラダイムとは異なり、IRL（Inverse Reinforcement Learning）（Adams et al., 2022）は観察された行動から報酬関数を導出することに焦点を当てる（Ng et al., 2000; Arora and Doshi, 2021）。標準的な IRL 手法には、最適な専門家の行動や意思決定過程を仮定する特徴マッチング法（the feature matching methods）（Abbeel and Ng, 2004）や、最適な行動を必要としない最大エントロピー法（the maximum entropy methods）（Ziebart et al., 2008）やベイズ法（the Bayesian methods ）（Ramachan- dran and Amir, 2007）がある。IRL は状態分布の変化に対する堅牢性をアシュアランスするが、余分な RL ステップによる計算複雑性の増大を代償とする（Ho and Ermon, 2016; Fu et al., 2018b）。一方、この相互作用は、サンプル効率（sample efficiency）（Yu, 2018）や環境との相互作用における潜在的な危険性（Garcıa and Fernández, 2015）など、RL 固有の課題をもたらす。さらに、報酬関数の特定は依然として課題である (Kim et al., 2021)。

2.3.2 Reinforcement Learning from Human Feedback (RLHF)【人間のフィードバックからの強化学習（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.

RLHF は DRL（深層強化学習）の領域で PbRL を拡張したもので（Christiano et al., 2017）、複雑な AI システムを人間の選好により近づけることを目的としている（OpenAI, 2023b）。RLHF の主な利点は、人間が、デモンストレーションを行ったり報酬を手動で設定したりするよりも、適切な行動を判断する方が得意であることを利用している点である。このアプローチは、特に LLM のファインチューニングにおいて大きな支持を得ている（Ouyang et al., 2022; OpenAI, 2023a; Touvron et al., 2023）。にもかかわらず、RLHF は、データの品質に関する懸念、報酬の誤汎化のリスク、報酬のハッキング、ポリシーの最適化における複雑さなどの障害に遭遇する（Casper et al., 2023b）。具体的には、RLHF は（§2.4.3 で示されるように）深層再帰的モデリングを伴わない再帰的報酬モデリング（Recursive Reward Modeling：RRM）プロセスとみなすこともできる（Leike et al., 2018）。ここでは、RLHF 手法の簡単なレビューを行う。

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.

RLHF の起源は Knox and Stone (2008, 2012) まで遡ることができ、その後、社会的ロボット (Knox et al., 2013) や人間-AI の協調学習（human-AI cooperative learning）（Griffith et al., 2013）などの領域にまでその範囲を広げている。Loftin et al.,（2016）は、フィードバックとポリシーの関連性に焦点を当てるだけでなく、フィードバックとトレーナーの戦略の関連性をモデル化している。Christiano et al., (2017) は、RLHF をシミュレーションされたロボットタスク（simulated robotic tasks）に拡張し、その潜在的な有効性を実証した。

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 (Garcıa 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.

RLHF の重要な応用例の一つが LLM の分野であることは注目に値する。RLHF を用いて訓練された LLM（Ouyang et al., 2022; Korbak et al., 2023; Christiano, 2023）は、教師あり学習や自己教師あり学習アプローチで訓練されたモデルと比較して、より創造的で人間のアラインメントが可能であること（Kenton and Toutanova, 2019; Brown et al., 2020b）を、いくつかの研究が発見した。RLHF の重要性は、単に LLM が人間の指示に従うことを可能にするだけにとどまらない（Ouyang et al., 2022）RLHF は、LLM に親切、無害、誠実といった重要な資質を与えることで、LLM をより良くアラインするのに役立つ（Bai et al., 2022a）。これらの改善により、多くの研究が LLM のアラインメントに RLHF を使用している（Ziegler et al., 2019; Stiennon et al., 2020; Bai et al., 2022a; Glaese et al., 2022; OpenAI, 2023a; Touvron et al., 2023）。さらに、Dai et al.（2024）は、Safe RL（Garcıa and Fernández, 2015）のフレームワークを RLHF と統合し、有用性と有害性の間の内在的な緊張に対処している（Bai et al., 2022a）。今後の取り組みとしては、人間によるアノテーションへの依存を減らすこと（Wang et al., 2023c; Sun et al., 2024）、反復的な RLHF 手法を活用することで報酬モデルの有効性を向上させること（例えば、ディベートフレームワークとの統合（Irving et al., 2018））などに焦点を当てることができる。Qiu et al.(2024) はまた、RLHF プロセスをテキスト分布上のオートエンコーディングプロセスとして描き、RLHF における収束特性（convergence properties）の分析を可能にする、RLHF プロセスの形式的フレームワークを構築した。

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:

Ziegler et al. (2019); Ouyang et al. (2022); Rafailov et al. (2024) の RLHF パイプラインをレビューし、一般的なフレームワークを示す。パイプラインは通常 3 つの段階からなる。

- **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.

- **教師ありファインチューニング（SFT）** RLHF は通常、事前に訓練された言語モデルから始まり、その後、教師あり学習、具体的には最尤推定を用いて、下流工程のタスクに合わせた高品質な人間の指示データセット上でファインチューニングを行い、モデル π^{SFT} を得る。これらのタスクの例としては、対話処理、指示追従、要約などがある（オープンソースのデータセットには、Alpaca Data (52k instruction-following data) (Taori et al., 2023)、Vicuna (70K user-shared ChatGPT conversations) (Chiang et al., 2023) などがある。この段階は、他のどの段階でも実施することができる。

- **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.

- **比較データの収集と報酬モデリング** この段階には比較データの収集が含まれ、このデータはその後報酬モデルの学習に使用される。SFT モデルには、$\pi^{\text{SFT}}(y \mid x)$ からサンプリングされた応答のペア (y_1, y_2) を生成するためのプロンプトが x として与えられる。これらのペアは、その後、人間のアノテーターに表示され、アノテーターは、いずれかの応答を選好することを示す。そして、§ 2.2 で議論したように、比較データは報酬モデル r_θ を構築するために使用される。

- **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:

- **強化学習によるポリシーの最適化** 最後のステップは、報酬モデル r_θ によって導かれる RL を通じて、LLM をポリシー π として最適化することである。プロンプトから LLM が応答を生成するプロセスは、次のようにモデル化される。RL はバンディット環境［損失を最小化する選択肢を選ぶディレンマ的環境］(Ouyang et al., 2022) において、各応答の最後に報酬モデル r_θ から報酬を得る。RL の主な目的は、訓練用プロンプトデータセット $\mathcal{D}_{\text{RL}}$ の期待報酬が最大になるように LLM のパラメータ ϕ を調整することである：

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

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:

通常、報酬の過剰最適化を緩和するために、SFT モデル π^{SFT} に由来するトークンごとの KL ペナルティが追加される。さらに、事前学習分布 $\mathcal{D}_{\text{pretrain}}$ からの勾配を統合することで、モデルの性能を維持することができ、Ouyang et al.(2022) では PTX 損失と呼ばれている。その結果、より包括的で実用的な目的関数が導入される：

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

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.

ここで、β と η はそれぞれ KL ペナルティの強さと訓練前の勾配の混合を決定する係数である。このプロセスにより、LLM は、訓練中に使用されたプロンプトに対する人間の選好によりアラインされた応答を生成するように改良される。

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.

RLHF は LLM を人間の選好に合わせるのに効果的であることが証明されているが、この方法には複雑な実装、ハイパーパラメータチューニング（hyper-parameter tuning）、サンプル効率（sample efficiency）（Choshen et al, 2019）、計算オーバーヘッド（computational overhead）（Yuan et al, 2024）などの問題があり、スケールアップが難しい。

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.

簡単なアプローチは、拒絶サンプリング（rejection sampling）(Dong et al., 2023; Touvron et al., 2023) と最良の例に対するファインチューニングの組み合わせである。すべてのプロンプトに対して、K 個の応答がモデルからサンプリングされる。各応答は報酬モデルで評価され、最も報酬が高いものが最良の応答として選択される。この選択された応答は、後にモデルのファインチューニングに使用される。

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.

Zhang et al.(2023a) は、言語モデル命令アラインメント問題（the language model instruction alignment problem）を目標到達型強化学習問題（goal-reaching reinforcement learning problem）として定式化し、HIR（Hybrid Iterative Reconstruction）アルゴリズムを提案している。この方法は、オンラインサンプリングとオフライントレーニングの 2 段階で展開される。オンラインサンプリングでは、アルゴリズムは LLM の温度パラメータ (temperature) を高くしてサンプリングする。オフライン学習段階では、生成された出力に基づ

いて命令を再ラベル化し、続いてこの再ラベル化されたデータを用いて教師あり学習を行う。HIR は、パラメータを追加することなく、成功したケースと失敗したケースを利用する。

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.

RRHF は、Yuan et al.（2024）によって紹介されたように、複数のソースからの応答をスコアリングしてランク付けすることによって、モデルの確率を人間の選好に合わせる。1 つか 2 つのモデルしか必要としないため、その実装は簡単である。RRHF は、言語モデルと人間の選好を効果的にアラインさせることができ、PPO と同等のパフォーマンスを生み出すことができると報告している。Gulcehre et al.,（2023）は、2 つのループを含む ReST アルゴリズムを提案している：成長（Grow）と改善（Improve）である。成長（Grow）ループは、現在のモデルを用いてデータセットをサンプリングして生成し、改善（Improve）ループは、固定されたデータセット上でモデルを反復的に学習する。このアルゴリズムは、計算効率を向上させるために固定データセットを繰り返し使用することを可能にする、シンプルで効率的なフレームワークを提供し、教師あり学習ベースライン［のモデル］と比較して、報酬モデルのスコアと翻訳品質の大幅な改善を示す。Chakraborty et al. (2024) は、RLHF におけるポリシーの最適化における報酬モデリングの優位性に動機付けられ、二段階最適化に基づくフレームワーク（bilevel optimization-based framework）である PARL を提案する。

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.

Rafailov et al.（2024）は、報酬関数と最適ポリシーの間のマッピングを示す DPO を導入している。DPO はシンプルかつ効率的であり、人間の選好データから直接言語モデルを最適化するため、明示的な報酬モデルや多段階学習が不要である。

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.

Azar et al(2023) は、ペアワイズ（pairwise）の人間の選好から学習するために設計された一般的な目的 ΨPO を提示し、現在の方法の仮定（ペアワイズ（pairwise）の選好をポイントワイズ（pointwise）での報酬で代替できるという仮定）を回避している。この目的は、RLHF と DPO の行動を分析し、潜在的な過学習の問題を明らかにする。著者らはさらに、Ψ を識別子（the Identity）として設定することで、ΨPO の特定のインスタンスを掘り下げ、過学習の問題を軽減することを目指している。彼らはこの方法を IPO と呼び、IPO と DPO を対比した実証結果を示している。Hejna et al(2024) は、最適ポリシーに関する情報を直接提供する、後悔に基づく選好モデル（regret-based model of preferences）を利用する CPL を導入している。

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.

さらに研究を進めることで、RLHF が LLM（大規模言語モデル）で効果的に機能する理由や、人間と AI の協働の利点を促進するために RLHF をマルチモーダルな設定（Yevgen Chebotar, 2023; OpenAI, 2023b）での適用を探ることができる（Carlson and Demiris, 2010; Wu et al., 2021; Bi et al., 2021）。RLHF の未解決の問題についてのサーベイを提供している Casper et al.(2023b) も参照。

2.4 Scalable Oversight【スケーラブルな監視】

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 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.

統計的学習アルゴリズムは通常、独立性や同一分布など、データ分布に関する特定の仮定に依存している。その結果、これらのアルゴリズムは、特に特定の分布の下では、いくつかの状況で失敗する (Zhou et al., 2022)。初歩的なシステムにおける課題は、目視検査によって速やかに特定することができる (Christiano et al., 2018; Ngo et al., 2024)。AI システムがより強力になるにつれて、学習信号の捕捉が不十分であったり、損失関数の設計を誤ったりすると、矛盾を難読化することで人間を欺く（Russell, 2019）、仕様ゲーミング（specification gaming）（Victoria et al., 2020）、報酬ハッキング（Brown et al., 2020a）、権力追求力学（power-seeking dynamics）（Carlsmith, 2022）といった破滅的な行動（Russell et al., 2015; Hubingeret al., 2019c; Cotra, 2021）につながることが多い。人間の視点から見ると、これらは AI システムの最適化された目標と、我々の頭の中にある理想的な目標との間にギャップがあることを意味する。したがって、様々な意思決定において効果的な監視を提供することが極めて重要になり（Bowman et al., 2022; Li et al., 2023a）、しばしば 2 つの現実的な課題から生じるスケーラブルな監視（Amodei et al., 2016）と呼ばれる。

- 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).
- AI システムの挙動を人間が頻繁に評価することによる高いコスト。例えば、訓練プロセスには時間がかかり、リアルタイムで人間の指示を訓練ループに組み込むことは、人的資源を著しく浪費し、訓練効率を阻害する（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).
- AI システムの動作は本質的に複雑であるため、特に理解しにくく、リスクの高いタスク（Saunders et al, 2022）、例えば、本を要約するように AI システムを教えるタスク（Wu et al, 2021）、複雑なコードを生成するタスク（Pearce et al., 2022）、将来の天気変化を予測するタスク（Bi et al, 2023）では、評価が困難になる。

Scalable oversight seeks to ensure that AI systems, even those surpassing human expertise, remain aligned with human intent.
スケーラブルな監視は、システムが人間の専門知識を凌駕するものであっても、人間の意図にアラインされたものであることを保証しようとするものである。

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).

この文脈において、我々の主な焦点は、スケーラブルな監視を構築するために、まだ一般的に実現されていないが、可能性のあるいくつかの有望な方向性を提示することである（Amodei et al, 2016; Leike et al, 2018）。

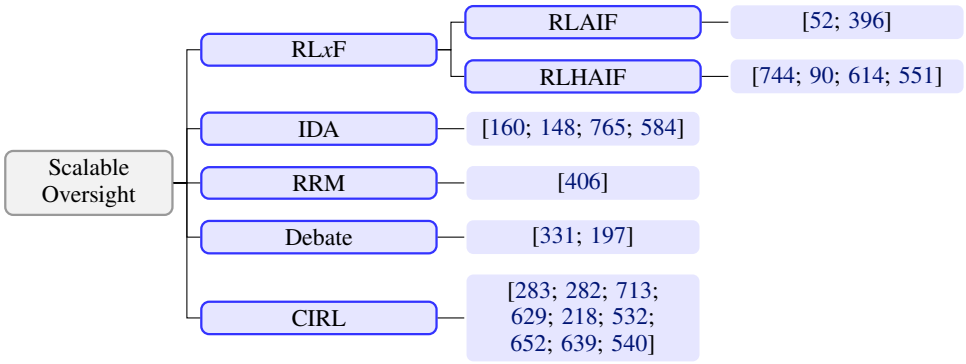

Figure 5: A tree diagram summarizing the key concepts and literature related to Scalable Oversight. The root node represents Scalable Oversight whose goal is *ensuring AI systems remain aligned with human intent even as they surpass human capabilities*. The main branches represent promising frameworks 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.
図５：スケーラブルな監視に関連する主要概念と文献をまとめたツリー図。ルートノードはスケーラブルな監視を表し、AI システムが人間の能力を超えたとしても、人間の意図にアラインされた状態を維持することを目標としている。メインブランチは、フィードバックからの強化学習（RLxF）、蒸留と増幅の反復（IDA）、再帰的報酬モデリング（RRM）、ディベート、協調的逆強化学習（CIRL）などの有望なフレームワークを表している。さらにサブブランチには、各フレームワークを探求する主要な研究がリストアップされている。この図は、AI システムの複雑化に伴い、効果的で安全な監視メカニズムを構築するための研究の方向性を概観するものである。

2.4.1 From RLHF to RLxF【RLHF から RLxF へ】

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.

RLHF パラダイムは、複雑なシステムをアラインするためのフレームワークを提供する（OpenAI, 2023a; Touvron et al., 2023）。しかし、人間による評価の不正確さや、それに伴う高いコストといった障害に遭遇する（Christiano et al., 2017; Casper et al., 2023b; Perez et al., 2023）。重要な限界は、超人的な能力を持つ AI システムを作成する際に、人間のフィードバックを拡張するために RLHF を利用することの難しさである（Wu et al., 2021）。RLHF パラダイムに基づき、我々はスケーラブルな監視のための基本的なフレームワークとして RLxF を導入し、フィードバックの効率と質を高め、より複雑なタスクのために人間のフィードバックを拡張することを目指す。これは、AI コンポーネントを組み込むことによって RLHF を強化するものである（Fernandes et al., 2023）。RLxF の x は AI と人間の融合を意味する。RLxF に関する具体的な方法論については、この後のセクションでさらに掘り下げていく。

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.

AI フィードバックからの強化学習（RLAIF） RLAIF は、RLHF のフレームワークを基に構築された手法であり、RLHF の拡張として機能する。Bai et al. (2022a) は、RLHF によって訓練された LLM は、しばしばデリケートで論争的な問題を回避することを選択し、モデルの全体的な有用性を低下させる可能性があることを発見した。これらの限界を考慮し、Bai et al.（2022b）は RLAIF に基づく学習パイプラインを提案した。このパイプラインは、人間のフィードバックではなく、LLM（例えば GPT-4 や超人的能力を持つ他の言語モデル）によって生成されたフィードバックを利用する。事前に設定された基準に基づいて、ポリシーモデルは自己評価し、レッドチームによって促された応答を修正する。そして最初のポリシーモデルを修正された応答でファインチューニングする。最後に、ファインチューニングされたポリシーモデルは、別の言語モデルの応答に対する有害性を評価する（すなわち、AI フィードバック）。RLHF 法をミラーリングし、彼らはこのフィードバックを用いて報酬モデルを訓練し、ポリシーモデルの振る舞いを最適化する。Lee et al. (2023a) は要約タスクにおいて、RLAIF と RLHF で訓練されたモデルのパフォーマンスの違いを比較している。彼らの結果は、AI フィードバックで訓練されたモデルは、人間によって評価されたとき、人間フィードバックで訓練されたモデルとほぼ同じ全体的なパフォーマンスを達成したことを示唆しているが、ニュアンスは異なる。

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).

ある程度、RLAIF は RLHF に内在する回避（evasion）（すなわち、実用性を著しく低下させることなく無害化を維持すること）に取り組んでいる。AI フィードバックは、最小限の人間の介入で訓練ループを構築し、訓練コストを削減するための実行可能な代替手段を提供する。透明でアクセス可能な AI 行動ガイドラインに従った AI 監視は、スケーラブルな監視の達成に大きく役立つ可能性がある (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.

人間と AI のフィードバックからの強化学習（RLHAIF） RLHAIF は人間と AI の要素を統合して監視を行う。Wu et al.(2021) は、書籍の要約において人間を支援する AI の実現可能性を調査している。この方法は、書籍要約タスクをサブタスクに分解してツリー状の構造を形成することで、人間による監視とモデルの性能評価を容易にした。一方、Saunders et al. (2022) は、モデルの有効性を人間が評価する際に AI を活用する可能性を調査している。その結果、モデルが生成する批評（critiques）は、人間が見逃しているかもしれない欠陥を特定するのに役立つことが示された。Bowman et al. (2022) は、サンドイッチング（Cotra, 2021）に基づくスケーラブルな監視技術を評価する有望性を実証するための POC（概念実証）実験を提案している。信頼性の低い LLM と協調した場合、人間はモデルや自分自身【信頼性の低い LLM と協調した人間】を有意に凌駕することが明らかになった。Perez et al. (2023) は、様々なスケールの言語モデルの動作を評価するためのデータセットを自律的に生成する言語モデルを採用している。著者らは、人間によって検証された 154 の高品質なデータセットを作成した。これらの方法は、複雑な問題や様々な領域において、人間の監視をスケールアップするために AI の支援を利用することの実現可能性を示している。

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.

考察　純粋な人間を他の構成要素に置き換えることで、RLHF アルゴリズムを強化する取り組みがいくつか進行中である（Leike et al, 2018）。人間のフィードバックが多次元的であることを考慮し、特定のルールに基づいた人間の判断を提供するための様々なアプローチが考案されている。そのようなルールの例には、チャットの流暢さ（Saunders et al.,2022）やプライバシー保護（Carr、2023）のような考慮事項が含まれる。Saunders et al.(2022) は、質の高い対話の要件を自然言語ガイドラインに分解し、エージェントがそれに従うべきであるとし、各ガイドラインごとに評価を求めている。我々は、対象となる人間の評価を収集し、このデータに基づいてモデルを訓練することにより、より効率的なルール条件付き報酬モデルを達成することができる。このアプローチは、プロンプト言語モデルと比較して、対話エージェントの有効性を大幅に向上させ、より有用で、正確で、良心的なものにする。Carr(2023) は、プライバシーの保護と言語モデルの出力品質の調和を目指す、プライバシーフィードバックからの強化学習 (RLPF) を提案している。この方法は、モデルによって生成されたテキストのリアルタイムのプライバシーリスク評価を行うために NLP（Natural Language Processing；自然言語処理）技術を利用し、その後、これらの評価に基づいて強化学習フィードバック信号を調整する。具体的には、生成されたテキストにセンシティブな情報が含まれている場合、ネガティブなフィードバックが発生し、一方、高品質で漏洩性のない（non-revelatory）テキストにはポジティブなフィードバックが発生する。モデルが訓練を受けるにつれて、その能力は漸進的に改善され、テキストの品質が向上し、プライバシーの侵害が同時に最小化される。このアプローチは、大規模な手動データアノテーションに大きく依存する従来の学習方法とは対照的に、確立された NLP 技術を採用することで、より効率的なプライバシーリスクの評価を提供する。

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.

RLxF 法の核心は、大きな問題を小さな下位問題に分解するという戦略を利用することで、AI やソフトウェアなどのより効率的なツールを使用して、下位問題を迅速に解決することを可能にしている。これらの下位問題の解決策を活用することで、主要な問題の解決を早めることができる。これらの技術は、IDA の初歩的な例とみなすことができる。主な違いは、継続的な反復プロセスがないことにある。それにもかかわらず、これらの手法は、人間のパフォーマンスを超える AI システムにフィードバックを提供する有望な手法であることが示唆されている（Wu et al., 2021）。その結果、これらの手法は、より高度な AI システムのトレーニングにおける基礎技術として役立つ。

2.4.2 Iterated Distillation and Amplification【蒸留と増幅の反復】

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.

蒸留と増幅の反復（Iterated Distillation and Amplification：IDA）は、人間と AI の反復的なコラボレーションによってスケーラブルな監視を構築するためのフレームワークを導入する（Christiano et al., 2018）。このプロセスは、人間 H の意思決定を反映した $A[0]$ と呼ばれる初期エージェントから始まる。 $A[0]$ は、人間に近いレベルの熟練度を備える強力な技術を用いた訓練を受ける（蒸留ステップ）；次に、H と複数の $A[0]$ インスタンスとの協調的相互作用により、強化されたエージェント $A[1]$ が作成される（増幅ステップ）。この連続プロセスはアルゴリズム 1 に記述されている。

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

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 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.
Cotra (2018) は、RL と IRL の両方において、広い定義と狭い定義を区別している。広義の RL は、AI システムに少ない（sparse）報酬シグナルを与え、将来の累積報酬の自律的な探索と最適化を可能にする。これは超人的な新規戦略（super-human novel strategies）を導くことができるが、何を気にかけるかを完璧に特定することが難しくなる。狭義の RL は、最終的な結果ではなく、選択の妥当性をフィードバックする。これは ML システムをより人間に近づけるが、能力は制限される。同様に、広義の IRL は人間のあらゆる行動から深い長期的価値を推論するが、狭義の IRL は短期的な手段的価値しか推論しない。前者はリスクが高く、後者は能力が制限される。

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 のトレーニングでは、各エージェントが人間の行動を模倣できるようにするために、狭い範囲のテクニックが必要となる。具体的には、狭義の RL や IL を使用して、エージェントを可能な限り人間のように制御できるように訓練することができる。人間は、エージェントの計算能力と並列処理能力を活用して、より先見性のあるマクロな戦略を考案することができる。これは本質的に、人間が本来持っている能力を増幅させるものである。次の反復では、エージェントはこの強化された人間と機械のシステムを再び模倣し、狭い範囲のテクニックを使用する。これにより、エージェントを人間的価値観にアラインさせながら、狭い能力から広い能力へと徐々に移行させることができる。反復が増えるにつれて、人間と機械のシステムはより高い能力を持つようになり、次第に高い能力を持ちながら人間的価値観に沿ったシステムに近づいていき、安全性と能力の両方を達成することができる。言い換えれば、狭いテクニックはエージェントが人間的価値観に従うようにするために使用され、増幅段階における人間の戦略の拡大はエージェントの活用方法であり、エージェント自身の学習目標を拡大するものではない。

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.
IDA は AlphaZero (Christiano et al., 2018; Nguyen, 2020) によってよく説明されている。このアルゴリズムは単純な方針（例えばランダムな打ち方（random move selection））から始まり、自己プレイゲームから学習する（増幅フェーズ）。そして、これらのゲームをトレーニングデータとして使用し、蒸留フェーズで、より優れた打ち方のヒューリスティック（better move selection heuristics）を開発する。この蒸留-増幅のプロセスを繰り返すことで、高速で熟達した碁を打つ AI を作ることができる。ここで、アラインメントとケイパビリティの区別が重要になる（Mennen, 2018）。アラインメントが取れているが能力（capable）の低い AI は勝とうとするが、中程度の相手には成功しないかもしれない。能力（capable）はあるがアラインメントが不十分な AI は、勝利以外の特定のゲーム特性を達成する。目標は、AI が能力（capable）を持ち、ゲームに習熟し、ゲームに勝つという目標に沿ったアラインメントをとることである。

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

Algorithm 1 Iterative Distillation and Amplification（蒸留と増幅の反復）

procedure IDA(H)
  $A \leftarrow$ random initialization（ランダム初期化）
  repeat
    $B \leftarrow$ AMPLIFY(H, A)
    $A \leftarrow$ DISTILL(B) ▷ Repeat indefinitely
  until False
end procedure
procedure DISTILL(overseer)
  return An AI trained using narrow, robust techniques to perform a task that the overseer already understands how to perform.
end procedure
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.
end procedure

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.

IDA の実現可能性については、かなりの議論が巻き起こっている（Yudkowsky, 2018）。IDA は、反復を通じてエラーが継続的に蓄積されることはないという重大な前提の下で動作する（Leike et al., 2018）。したがって、蒸留と増幅のステップでは技術的な課題が残り、十分に高度で安全な学習技術が必要となる。さらに、原著者（original authors）は IDA を AlphaZero（Silver et al., 2017）のトレーニングプロセスになぞらえ、簡易的試験環境（toy environments）で実証したにもかかわらず（Christiano et al., 2018）、その実用性は、H が複雑なタスクの一部を A に委譲することができるかどうかにかかっている。実際、Gato（Reed et al., 2022）は、AGI への道を開くかもしれない IDA（Mukobi, 2022）の重要な側面を示している。複数の専門家 AI の能力を 1 つのモデルに統合し、IDA の蒸留が現代の深層学習を使って達成できることを検証している。完全には実現されていないが、Gato は増幅の可能性を示唆しており、多様なスキルを活用して新しいタスクの学習を加速させる。しかし、Gato はアラインメント特性を維持するための安全な増幅や蒸留の方法を欠いている。Gato のようなモデルに適した、アラインメントを維持する IDA 手法を構築することは、AI の安全性研究にとって極めて重要な方向性である。要するに、Gato は IDA の実用化において注目すべき進歩を示しているが、IDA フレームワークが安全な AGI につながることを保証するためには、さらなる理論的進歩が不可欠である。

2.4.3 Recursive Reward Modeling【再帰的報酬モデリング】

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.

2.2 節で述べたように、報酬モデリングは、人間のフィードバックを使って報酬モデルを訓練し、それをエージェントが追求するという考え方を活用する。これにより、システムの目的の構築とその行動の評価を切り離すことができる（Ibarz et al., 2018）。このように、報酬モデルは AI システムの最適化の方向性に対する洞察を提供する。特に注目すべきは、人間の指示に従うように言語モデルをファインチューニングするなど、システムを人間の意図や価値観にきめ細かく合わせる能力である（Bai et al., 2022a; Touvron et al., 2023）。また、報酬モデリングは、AI 研究を進める上で有用であることが証明されている（Zhao et al., 2023; Bukharin et al., 2023）。再帰的報酬モデリング（RRM）（Leike et al., 2018）は、報酬モデリングの応用をより複雑なタスクにまで広げようとしている。複雑なタスク RRM の中心的な知見は、すでに訓練されたエージェント A_{t-1} を再帰的に使用し、より複雑なタスクで後続のエージェント A_t の訓練のために、それ自身の増幅されたバージョンで報酬学習を実行することによってフィードバックを提供することである。A_0 は、基本的な報酬モデリング（純粋な人間のフィードバックから学習）によって訓練される。このアプローチは、人間のフィードバックに影響されるだけでなく、やりがいのある結果を構成するものについてのモデル自身の評価にも影響される。結果を評価することは、行動を生み出すことよりも簡単であるという仮定が成り立つならば、報酬モデリングの反復プロセスは、より強力な AI システムを監督するためのより高い能力を反復的に達成することができ、監督をより複雑な領域に拡張する道を開く。このプロセスはアルゴリズム 2 に詳述されている。

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.
例えば、私たちは AI A に総合的な都市計画を立案させることを目指している。都市の設計には、交通計画、公共施設、住宅地と商業地の分布など、多くの複雑な要素が含まれる。都市全体の設計を評価することは重要な課題である。なぜなら、多くの問題は、実世界で長期間テストして初めて明らかになるからである。このプロセスを支援するために、交通計画に特化したエージェント B が必要になるかもしれない。しかし、交通計画はそれ自体が多面的なタスクである。そのため、道路幅、交通の流れ、公共交通機関の設計などの側面を評価する他のエージェントがさらに必要になる。道路幅の測定などのサブタスクごとに、安全基準が満たされているか、様々な交通手段が考慮されているかなどを検証する補助エージェントを訓練することができる。このようにして、各エージェントがサブタスクを評価するエージェントの助けを借りて訓練される RRM プロセスを確立する。

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.
このアプローチは、大企業の組織構造に似ている（Leike et al, 2018）。都市計画の文脈では、主要な計画チーム（CEO）が最終的な設計決定に責任を負う。彼らの決定は、交通チーム（部長）からの推奨（recommendations）に基づき、さらに交通チームは道路幅員チーム（管理職）からのインプットに基づき、推奨を行うといった具合である。意思決定の各レベルは、その下のレベルからのフィードバックに依存し、各タスクは報酬モデルによって最適化される。

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).

Algorithm 2 Recursive Reward Modeling(再帰的報酬モデリング)

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**

RRM が直面する課題は、アウターアラインメントとインナーアラインメントという概念を中心に説明することができる（Hubinger, 2020）。アウターアラインメントとは、学習された報酬モデルが、行動モデルによって認識された領域において正確な分布であることを保証するためのフィードバックメカニズムの充足性を中心に展開される。この課題は、人間のフィードバックの質、汎化の難しさ、エージェントの欺瞞の可能性など、いくつかの要因に左右される。逆に、インナーアラインメントは、報酬モデルとエージェントの両方における欺瞞的な行動や破滅的な行動を防ぐために、人間が透明性ツールをいかに効果的に用いることができるかに集中する。これは、監視メカニズムの有効性と、報酬モデルが最適化されておらず、エージェントが短期的視点にとどまっていることを検証する能力にかかっている（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.

これらの課題を軽減するための潜在的なアプローチ (Leike et al., 2018) には、訓練中に報酬モデルを修正するためのオンラインフィードバック (Christiano et al., 2017)、安全でない状態について教えるためのオフポリシーフィードバック（off-policy feedback）(Everitt et al., 2017)、教師なし学習やアノテーションによる動画やテキストなどの既存データの活用 (Baker et al、2022)、異なるレベルでの階層的フィードバック（Bukharin et al., 2023)、脆弱性を発見するための敵対的訓練（Madry et al., 2018)、フィードバックを募るための不確実性推定（Hadfield-Menell et al., 2016; MacGlashan et al., 2017）などがある。RRM の強みはその競争的トレーニングアプローチであり、実演の代わりに人間によるフィードバックが必要であり、フィードバックをより信頼性の高い、よりシンプルなものにする可能性がある (Hubinger, 2020)。要するに、RRM のプロセスは IDA(Christiano et al., 2018) になぞらえることができ、そこでは教師あり学習や模倣学習の代わりに報酬モデリングが行われる。したがって、RRM が直面する課題は、IDA で遭遇する課題、特にエラーの蓄積を防止する上で遭遇する課題を忠実に反映している。さらに、報酬モデリング自体は、狭いモデルを抽出する必要はなく（Cotra, 2018)、アラインメントの程度とパフォーマンスのトレードオフに課題を提示する。

2.4.4 Debate【ディベート】

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.

ディベートは、アルゴリズム 3 で定義されるように、意思決定における人間の判断 (Irving et al., 2018) を支援するために、2 つのエージェントが回答と言明（answers and statements）を提示する。これは、エージェントが人間の審判からより高い信頼を得るために努力しながら、お互いの欠点を特定しようとするゼロサムディベートゲームであり、スケーラブルな監視を構築するための潜在的なアプローチとなり得る。例えば、囲碁のゲームでは、人間の審判は一枚の碁盤の有利な面そのものを見極めることはできないかもしれない。しかし、対局の過程と最終的な結果を観察することで、これらの審判はより容易にそれを推理することができる。

The premise of this method relies on a critical assumption: *arguing for truth is generally easier than for falsehood*, 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).

この方法の前提は、「真実を論証することは、嘘（falsehood）を論証することよりも一般的に容易であり、真実を論証する側に有利である」という重大な仮定に依拠している。しかし、この仮定は万能ではない。例えば、複雑な問題では、人間は討論で使われる専門的な概念を理解できないかもしれない。さらに、勾配降下の限定的な性質が、望ましくない循環パターン（すなわち、正直さや欠点の強調など、ある性質を最適化するとき、モデルはしばしば別の性質を見落としたり、低下させたりする）をもたらすかもしれない（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.

LLM の能力が進歩したことで、すでにディベートの実践例が見られるようになったことは特筆に値する（Du et al., 2023; Claude, 2023）。現実世界の特定のシナリオでは、ディベートに課題が生じる可能性がある（Irving et al., 2018）。例えば、ある種の質問は人間の理解には複雑すぎたり、全体像を提示するには量が多すぎたりすることがある。10 ギガピクセルの画像を解釈する複雑さや、インターネット全体の広大さをふるいにかけることを考えてみよう。同様に、質問に対する最適な回答が非常に長くなる場合もある。100 ページにわたる回答が必要な場合を想定してみよう。これらをナビゲートするために、エージェントは最初に回答を選択し、議論が進むにつれて、質問と回答のどちらかのセクションを明らかにするかもしれない。Irving et al.（2018）は、このプロセスに関する簡易的な実験（toy experiment）を行っている。一方で、人間の時間の限界を認めなければならない。ロボットに指示を出すなど、環境との相互作用が必要なシナリオでは、それぞれの行動が明確な議論を要求するかもしれない。時間的な制約から、人間がすべての議論を判断することは必ずしも可能ではない。この課題に対して、人間のフィードバックを予測する ML モデルを設計する必要があるかもしれない。

Another consideration is the convergence of the debate mechanism (Irving et al., 2018). Du et al. (2023) showcases 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.

もう 1 つの考慮点は、ディベートのメカニズムの収束である（Irving et al., 2018）。Du et al. (2023) は、たとえ正確性が保証されなくても、最終的には特異な回答に収束するディベートのフレームワーク固有の傾向を示している。一方、収束の達成に課題が生じた場合、収束の有効性を測るために直感に頼らざるを得ないかもしれない。このことは、人間の評価者の介入が必要であることを意味し、このような人間の評価者には一定レベルの専門知識が要求され、対処すべき課題が提起される。

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.

Algorithm 3 Debate（ディベート）

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

さらに、多様な視点から発信される多くの議論がある。Ngo (2021) は、ディベートは反復増幅の一種であり、具体的な研究課題に対して安全性を確保するためのものとして、より具体的に位置づけており、その敵対的なフレーミングは問題点を発見しやすくする、と考えている。Michaelcohen（2020）は、判定プロセスを揺さぶることを目的とした欺瞞的な戦略を採用するようディベーターに動機を与えることの悪影響に関する懸念を表明している。Armstrong (2019); Barnes (2020) は、難読化された議論問題（the obfuscated arguments problem）、曖昧な回答、誤解を招く含意の伝播といった課題を含め、ディベートプロセスに浸透し得る様々な問題について解説している。別の論者は、論証の根底に欠陥が存在する可能性は十分に低いと断言し、信頼性を主張することができるが、反対の論者は、論証のどこかに欠陥が存在する可能性は十分に高いと断言し、信頼性の欠如を主張することができる。

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.

Beth Barnes（2020）は、討論者がより有益な回答をするよう動機付けるために、反対尋問の概念を導入している。このプロセスでは、討論者は事前に精査すべき主張を選択し、相手側討論者の回答のコピーを入手することができる。このやりとりはすべて文書化され、討論者は関連する部分をジャッジに提示することができる。反対尋問の導入は、不誠実な討論者がジャッジを欺くために、事前の主張とは対照的な大げさな物語を利用することに対する強固な抑止力となる。

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.

ディベート（Irving et al., 2018）、IDA（Christiano et al., 2018）、RRM（Leike et al., 2018）の間には顕著な類似性が存在する。これらのアプローチは、「評価は、タスクを完了させることよりも簡単である」という基本原則から理解することができる。したがって、AI システムの評価能力を活用することで、人間にとってより有利な能力分布をもたらすことができる。これらの方法が直面する課題、特にエラーの蓄積を軽減する上での課題も類似している。

2.4.5 Cooperative Inverse Reinforcement Learning【協調的逆強化学習：CIRL】

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

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

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.

これまでのほとんどすべての手法は、フィードバックからの学習を推論や制御とは別のプロセスだと考えており、フィードバック提供者を環境の外部に存在する存在として暗黙のうちに考えていることが多い。実際、操作 (Shevlane et al., 2023) や報酬改ざん (Everitt et al., 2021) のような失敗モードは、環境の外部にあるはずのフィードバックメカニズムが環境の一部となり、AI システムの影響を受けるようになったときに発生する。しかし、協調的逆強化学習（CIRL）のフレームワークは、制御とフィードバックからの学習を統合し、人間のフィードバック提供者を同じ環境内の同僚のエージェントとしてモデル化する。CIRL は、監視を強化するのではなく、AI システムが監視をゲーム化するインセンティブを排除し、フィードバックを与える人間と AI システムを敵対的ではなく協力的な立場に置くことで、スケーラブルな監視問題にアプローチする（Shah et al., 2020）。CIRL のパラダイムでは、AI システムは人間の選好に合わせて一方的に最適化するのではなく、人間の真の目標を達成するために人間と協力する。

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 の動機と一般的な考え方 例えば、報酬ハッキング（Victoria et al., 2020; Skalse et al., 2022）、欺瞞（Park et al., 2023b）、操作（Shevlane et al., 2023）を含む多くのミスアラインメントの態様は、AI システムが誤って指定された目的に対して自信を持って最適化した結果である（Pan et al., 2021）。訓練とデプロイの間、指定された目的（例えば報酬関数）は AI システムにとって挑戦不可能な真実の役割を果たし、人間のフィードバックは目的に指定された範囲でのみ尊重されるため、改ざんされたり（Everitt et al., 2021）、操作されたり（Shevlane et al., 2023）する可能性がある。

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).

CIRL (Hadfield-Menell et al., 2016, 2017b; Shah et al., 2020) は、(1)AI システムに報酬関数に関する不確実性を明示的に保持させ、(2) 報酬関数の真意に関する唯一の情報を人間に提供させることで、この問題の軽減を試みている。この不確実性により、AI システムは人間に従う傾向を持ち、人間が本当に望んでいることを判断しようとする。具体的に言えば、人間のプレーヤー H とロボットのプレーヤー R が共通の報酬関数 $r(\cdot)$ を共有する 2 人協力ゲームとしてタスク全体をモデル化する。重要なことは、報酬関数と報酬信号は R には見えず（実際、トレーニングメカニズムによって明示的に計算されることはなく）、IRL（逆強化学習）に似たプロセスを通じて、R が H の行動から推測する（H に質問し、やり取りすることを含む）。このゲームは、CIRL（Hadfield-Menell et al., 2016）、アシスタンスゲーム（Fickinger et al., 2020）、およびアシスタンス POMDP（Shah et al., 2020）と呼ばれる。

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.

要するに、AI システムは人間の真の目的 $r(\cdot)$ を自らの目標としており（$r(\cdot)$ の値を確実に知っているわけではないにもかかわらず）、人間を観察し、人間と対話することによって r を常に把握しようとする。これにより、以下のようなインセンティブが減少する。というのも、例えば、人間の行動を操作することは、情報源を汚染するだけで、r には影響しないからである。

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.

CIRL の定式化 Hadfield-Menell et al.（2016）は、古典的なマルチエージェント MDP をベースに CIRL の設定（M とする）を特徴付け、M を以下のように定義している。

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

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 $\boldsymbol{\theta}$; $r : S \times \mathcal{A}^{\mathbf{H}} \times \mathcal{A}^{\mathbf{R}} \times \Theta \to \mathbb{R}$ 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).

上式において、S と $\{\mathcal{A}^{\mathbf{H}}, \mathcal{A}^{\mathbf{R}}\}$ はそれぞれ世界の状態とアクションの空間であり、$T : S \times \mathcal{A}^{\mathbf{H}} \times \mathcal{A}^{\mathbf{R}} \to \Delta(S)$ は推移関数、γ は割引率である。ここまでは、標準的なマルチエージェント MDP の定義と同じである。報酬関数はパラメータ化され、そのパラメータは分布でモデル化できる。Θ はパラメータ $\boldsymbol{\theta}$ の値の空間であり、$r : S \times \mathcal{A}^{\mathbf{H}} \times \mathcal{A}^{\mathbf{R}} \times \Theta \to \mathbb{R}$ は共有報酬関数であり、$P_0 \in \Delta(S \times \Theta)$ は初期状態と報酬関数のパラメータの結合分布である。このパラメータ化アプローチにより、R は真の報酬関数に対する信念を明示的にモデル化し、推論することができる。Nayyar et al. (2013) のテクニックを使用すると、任意の CIRL の設定は、等価な単一エージェント POMDP に還元することができ、その結果、比較的扱いやすい最適ポリシーの存在を証明することができる（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).

CIRL 研究の注目すべき方向性 H が R を教えることの重要性を積極的に強調する研究もあるが（Fisac et al., 2020）、人間が常に最適な共同のポリシー（optimal joint policies）に基づいて行動するという仮定は非現実的なものであるため、ゲーム均衡や共同のポリシー（H の教育行動を含む）の重視に異議を唱え、代わりに H のポリシーに対する R の最適応答に注目する研究 (Shah et al., 2020) もある。より具体的には、Shah et al.(2020) は、R の任意のポリシーに対する H のポリシー応答の分布を指定するポリシー条件付き信念 $B : \Pi^{\mathbf{R}} \to \Delta\left(\Pi^{\mathbf{H}}\right)$ を考慮し、目的は B を与えられた R の最適ポリシーを見つけることである。ここで、B は本質的に人間モデリングの一形態であり、1 つの課題は B として強力に正確な人間モデルを得ることである (Hong et al., 2022)。

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.

別の面では、Hadfield-Menell et al.（2017b）と He and Dragan（2021）は、H が真の報酬関数に関する情報を伝達する方法として、不完全な報酬関数を手動で指定することを検討している。これには、R 側の作業（すなわち、R が不完全な仕様に基づいて真の報酬関数に関する推論を実行できるようにすること）（Hadfield-Menell et al.,2017b）と、H 側の作業（すなわち、真の報酬関数をよりよく伝えるより堅牢な仕様を H が作成するのを支援するアルゴリズムツールの開発）（He and Dragan, 2021）が含まれる。ゲーム設定の改善もさることながら、よりスケーラブルな CIRL アルゴリズムの設計も優先課題として認識されている。

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.

また、CIRL とアシスタントゲームを、ロボットがサービスを提供する必要のある複数の人間が存在するマルチエージェント設定 (Fickinger et al., 2020) に拡張する研究もある。これは Critch and Krueger (2020) の

マルチ／シングル委任設定に相当し、人間の目的が様々であることが課題となり、社会的選択の手法の使用が必要となる。

2.5 Weak-to-Strong Generalization【弱から強への汎化】

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.

スケーラブルな監視は、より賢く複雑な AI システムに対して人間が監督信号（supervision signals）を提供し、超人レベルの AI システムの行動が人間の意図や価値観にアラインされたものであることを保証するのに役立つ。しかし、スケーラブルな監視の信号が得られない場合はどうだろうか？例えば、タスクによっては、評価が生成よりも単純であるとは限らず、スケーラブルな監視を実現するために、タスクの分解と AI による支援を利用することは不可能である。

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.

最近、弱から強への汎化（Weak-to-Strong Generalization）と呼ばれる汎化現象が検証されているが、その核となる考え方は、弱いモデルからの弱い監督信号を用いて強いモデルを訓練することである（Burns et al., 2023）具体的には、弱いモデルを実際のデータ（ground truth）で学習させ、新しいデータに弱いラベルをアノテーションして強いモデルを学習させる。3 つの設定（すなわち、NLP 分類、チェスパズル、報酬モデリング）にわたる結果は、弱から強への汎化が堅牢な現象であることを反映しているが、弱いラベルで訓練された強モデルと実際のデータ（ground truth）との間のギャップを狭めるなど、さらなる改善の余地がある。弱から強への汎化は、スーパーアラインメント問題、つまり人間がスーパー AI システムを弱い監督者としてどのように監督できるかという問題に、貴重なアナロジーを提供する。弱いものから強いものへの汎化の背後にある知見は、強いモデルが単に弱いモデルの動作を模倣するのではなく、弱いラベルを超えて汎化できるということである。言い換えれば、弱いモデルが強いモデルの能力を引き出すということである。しかし、人間が実際のデータ（ground truth）を知らなければ、弱から強への汎化を検証することは困難である。それでもなお、弱から強への汎化は、スーパーアラインメント問題を解決するための貴重な視点を提供する。

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.

弱から強への汎化のフレームワークは、スケーラブルな監視によってさらに拡張され、統合されつつある。経験的な結果は、弱いモデルが、2 つの専門家モデル間のディベートを評価することで、強いモデルの正しさを評価できることを示している（Khan et al., 2024）。さらに、専門家の討論をより説得力のあるものにすることで、非専門家のディベートにおける真実を見極める能力が向上し、実際のデータ（ground truth）なしでモデルをディベート戦略に合わせることの有効性が証明される。いくつかのフレームワークでは、蒸留と増幅を反復する外部増幅器を採用しており、学習プロセスにおいて、弱-強汎化技術を IDA と統合するための潜在的なフレームワークを提示している（Ji et al., 2024a）。さらに、Leike (2023a) は、スケーラブル

な監視と**弱から強への汎化**技術を統合するためのいくつかの方法を提案している。例えば、タスクを（スケーラブルな監視の原則に沿って）個々の（atomic）タスクに再帰的に分解し、これらの個々のタスクを監督し、人間の選好データを用いて**弱から強への汎化**技術で訓練された報酬モデルを採用する。

3 Learning under Distribution Shift【分布シフト下での学習】

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).

信頼性の高い AI システムの構築は、多様なデータ分布への適応能力に大きく依存する。訓練データと訓練環境は、実際のデプロイシナリオを不完全に近似したものであることが多く、以下のような重要な要素が欠けていることがある。 対抗的な圧力（adversarial pressures）（Poursaeed et al., 2021）（例えば、教師あり学習ベースのシステムにおけるガウスノイズ（Gilmer et al., 2019）や自動運転システムにおけるシャドウ攻撃（Ma et al., 2012））、マルチエージェントの相互作用（Critch and Krueger, 2020; Dafoe et al., 2021）、人間の監督者が効率的に評価できない複雑なタスク（Leike et al., 2018）、およびゲーム化や操作が可能な報酬メカニズム（Krueger et al., 2020）などである。このような訓練分布とテスト分布（または環境）の不一致は、分布シフトとして知られている（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.

したがって、トレーニング分布の下ではアラインメントがとれている（すなわち、人間の意図に沿った目標を追求している）AI システムでも、デプロイメント（またはテスト）分布の下ではアラインメントが維持されず、デプロイメント後に深刻なミスアラインメントの問題を引き起こす可能性がある。この潜在的な失敗が、データ分布にまたがるアラインメント特性（すなわち、人間の意図や価値観への準拠）の維持に関する研究の動機付けとなっている。

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

アラインメントの観点からは、目標を追求する能力の欠如とは対照的に、アラインメントが取れていない有害な目標を追求する AI システムをより懸念している。したがって、アラインメント特性に重点を置くということは、能力の汎化とは対照的に、分布全体にわたる目標の汎化に重点を置くことを意味する（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.

本節では主に、分布シフトの下で学習する際のアラインメント特性の保存について議論する。まず、分布シフトによるアラインメントの課題を紹介する（§3.1）。続いて、分布シフトに対処するための方法を掘り下げ、特に 2 つのアプローチについて議論する：(1) 学習プロセス中に最適化を誘導するアルゴリズム的介入 (§3.2) と、(2) 敵対的学習 (Yoo and Qi, 2021; Bai et al., 2021; Ziegler et al., 2022) や協調的学習 (Dafoe et al., 2021) を含む学習プロセスに特定の要素を導入することで学習分布を拡張するデータ分布介入 (§3.3) である (§3.3.2)。分布シフト下での学習の枠組みを図 6 に示す。

[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.

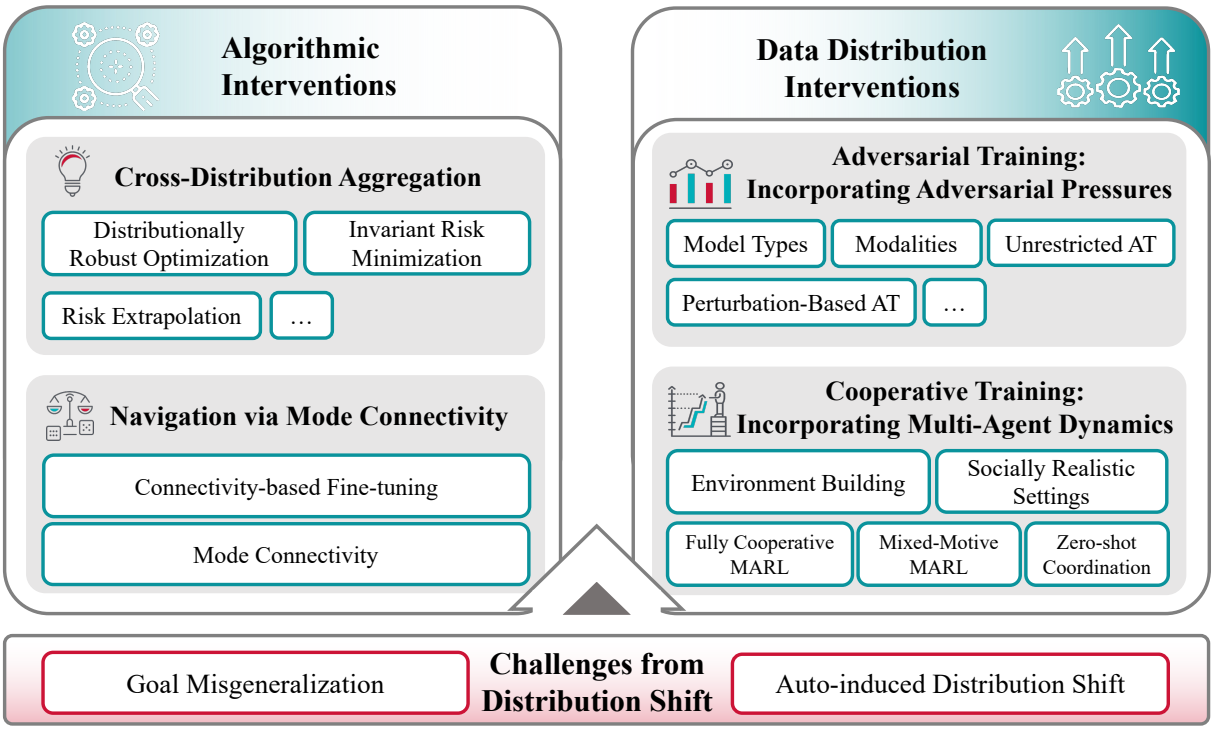

Figure 6: Framework of learning under distribution shift. The main challenges stemming from the distribution shift are goal misgeneralization and auto-induced distribution shift (§3.1). In our framework, we also introduce two kinds of methods to address distribution shift: algorithmic interventions (§3.2) 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.
図 6：分布シフトの下での学習のフレームワーク。分布シフトに起因する主な課題は、目標の誤汎化と自動誘発分布シフトである（§3.1）。また、本フレームワークでは、分布シフトに対処するための 2 種類の手法を紹介する。すなわち、学習中に最適化を誘導するアルゴリズム的介入（§3.2）と、実世界の要素を導入することで学習分布を狙い通りに拡張するデータ分布介入（§3.3）である。

3.1 The Distribution Shift Challenge【分布シフトの課題】

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).
具体的な手法を紹介する前に、まずアラインメントにおける主要な課題の 1 つが、なぜ分布シフト下での学習なのか、より具体的には、分布シフト下でのアラインメント特性（すなわち、人間の意図や価値観への忠実さ）の保持なのかを示す。分布シフトの問題に関連する 2 つのアラインメントの課題、すなわち、目標の誤汎化（Di Langosco et al.、2022）と自動誘発分布シフト（ADS）（Krueger et al.、2020）を紹介する。

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).
AI システムの訓練は、訓練入力分布の下での訓練報酬／損失の追求に従うように最適化される。しかし、この遵守は、入力分布が質的に変化する場合、すなわち分布がシフトする場合には、汎化されない可能性がある。このような変化には、例えば、敵対的圧力（Poursaeed et al.、2021）、マルチエージェント相互作用（Critch and Krueger、2020）、人間の監視者が効率的に評価できない複雑なタスク（Di Langosco et al., 2022）、ゲームや操作が可能な報酬メカニズム（Krueger et al., 2020）などが含まれる。

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.
ここでは、2 つの異なる失敗モードを区別する価値がある：目標の誤汎化（Di Langosco et al, 2022）は、元の分布とシフトされた分布が与えられ、自動誘導分布シフト（Krueger et al, 2020）は、AI システムが報酬を追求するために自身の行動でデータ分布を変更する。

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.
目標の誤汎化 この種の課題は、AI システムがトレーニング分布では完璧に機能するが、トレーニング分布で学習した能力が OOD（out-of-distribution；分布外）デプロイで汎化できず、AI が人間の希望に沿わない目標を追求するシナリオを指す（Di Langosco et al., 2022）。目標の誤汎化は、エージェントが OOD 環境において無能になる他の形態の誤汎化（能力の誤汎化など）とは区別される； その一方で、目標の誤汎化を抱えるエージェントは、OOD 設定において望まない目標を有能に追求する。

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).
単純化された例として、疑似相関（spurious correlations）（またはショートカット特徴）の事例がある（Geirhos et al, 2019; Di Langosco et al, 2022）。例えば、画像分類データセットにおいて、緑の草は牛というラベルを予測する高い特徴である。しかし、この特徴は様々なデータ分布において、より一貫性があり信頼できる必要があることに注意する必要がある（Murphy, 2023）。さらに、IL における因果的混乱（the causal confusion）（すなわち、助言者と環境との間の相互作用の因果構造を無視すること）は、目標の誤汎化をもたらす可能性がある（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.
目標の誤汎化がもたらす大きな危険の 1 つは、「人間が本当に望んでいることに最適化すること」と「人間が賛同すること（thumbs-ups）に最適化すること」の区別がつかないこと（indistinguishability）にある。しかし、AI システムは人間の選好に意図的に従ったり、人間から高い報酬を得るために欺いたりすることがあるが、実際には意図した目標、すなわち人間が本当に望んでいることを学習していない。後者には、人間の評価者を欺いたり、人間が賛同する（thumbs-ups）ように操作したりする可能性が含まれる (Shevlane et al., 2023)。例えば、Amodei et al. (2017) は、ロボットハンドが小さなボールをつかむタスクにおいて、ロボットハンドがレンズの前で視差を利用することで、実際にはボールをつかまずに、あたかもボールをつかんだように見せかける動作を発見した。この動作により、人間のアノテーターはタスクが完了したと勘違いする。

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).

[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).

AI システムを人間のフィードバックで訓練またはファインチューニングする場合、訓練ではどちらも完璧に機能するため、2 つの目標を区別することは不可能であり、AI システムがどちらを学習するかは不明である。実際、訓練中に人間の評価者が騙されたり操作されたりする可能性があり、AI システムは、人間が望むことよりもむしろ、人間の賛同（thumbs-ups）に最適化するよう、より強く動機付けられる可能性がある。この現象の現在の例は、レコメンダシステム (Kalimeris et al., 2021; Adomavicius et al., 2022)、LLM(Perez et al., 2023)、RL システム (Amodei et al., 2017) に存在する。

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.
最後に、目標の誤汎化と密接に関連する失敗モードの一つは、メサ最適化（Hubinger et al., 2019c）のミスアラインメントであり、学習されたモデル重みを持つ ML モデルが推論中にそれ自身の中で最適化（「メサ最適化」）を行い（Hubinger et al., 2019c; Dai et al., 2023）、この最適化の目的がモデルの学習目的にアラインされていない。

Auto-Induced Distribution Shift (ADS)【自動誘発分布シフト】 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.
AI システムをトレーニングする際、我々はエージェント自身の長所と短所のみを考慮し、これらのエージェントが環境に与える影響を見落とすことが多い。これまでの研究では、データが独立かつ同一に分布していると仮定することが多く（Besbes et al., 2022）、アルゴリズムがデータ分布に与える影響を無視してきた。しかし、Krueger et al. (2020) は、現実には、エージェントは意思決定と実行の過程で環境に影響を与え、その結果、環境によって生成されるデータの分布を変化させる可能性があると仮定した。彼らはこの種の問題を ADS と呼んでいる。

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.
実際の例としては、レコメンダシステムにおいて、レコメンダアルゴリズムによって選択されたコンテンツがユーザーの選好や行動を変化させ、ユーザー分布のシフトにつながる可能性がある。分布シフトは、今度はレコメンダアルゴリズムの出力にさらなる影響を与える（Carroll et al.、2022）。AI システムがますます世界に影響を与えるようになるにつれ、エージェントが人間社会に統合された後、社会全体のデータ分布にさらなる影響を与える可能性も考慮する必要がある。

3.2 Algorithmic Interventions【アルゴリズム介入】

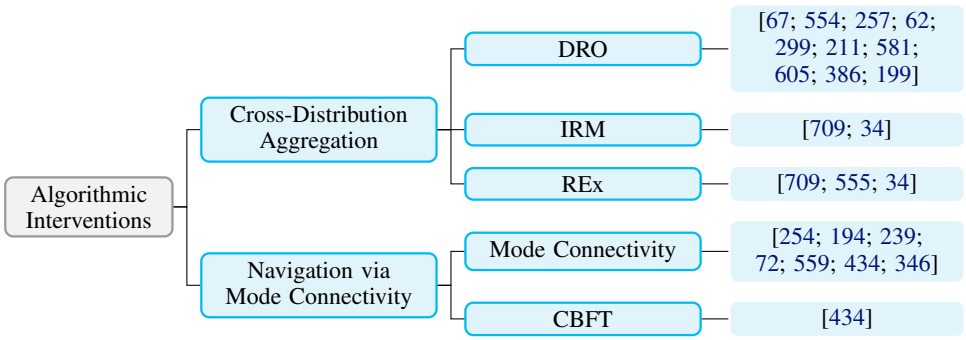

Figure 7: A tree diagram summarizing the key concepts and literature related to Algorithmic Interventions. The root node represents Algorithmic Interventions that aim to steer optimization during the training process. The main branches represent two main methods, namely cross-distribution aggregation (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).
図 7：アルゴリズム介入に関する主要概念と文献をまとめたツリー図。ルート・ノードは、訓練プロセス中の最適化を目指すアルゴリズム的介入を表す。メインブランチは、2 つの主な手法、すなわち、クロス分布集約（疑似特徴の代わりに不変関係に基づく予測子を見つけるために、訓練中に異なる分布のリスクを最小化することを目的とする）と、モード接続性を介したナビゲーション（モデルの汎化性能を向上させるために、モード接続性に基づくファインチューニングを目的とする）を表す。さらなるサブブランチは、分布ロバスト最適化（DRO）、不変リスク最小化（IRM）、リスク外挿（REx）、接続性に基づくファインチューニング（CBFT）などの重要な技術をリストアップしている。

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).
アルゴリズム介入の方法を説明するに際し、まず、分布シフトを緩和するために、トレーニング中に様々な分布上のオプティマイゼーションを誘導する 2 つのクラスの方法、すなわち、クロス分布アグリゲーション（aggregation：集合体、集約）（§ 3.2.1）とモード接続性（mode connectivity）を介したナビゲーション（§ 3.2.2）について概説する。

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).
第 1 部では、経験的リスク最小化（ERM）（Vapnik, 1991）の初期アプローチから、モデルの疑似特徴（spurious features）への依存から生じる問題を軽減するために考案された手法であるリスク外挿（REx）（Krueger et al., 2021）までの手法を取り上げる。第 2 部では、接続性に基づくファインチューニングを紹介する。これは、疑似でない相関関係への収束を促すために、学習中の損失ランドスケープのナビゲーションをガイドするもので、モード接続性からの知見を利用している（Lubana et al., 2023）。

3.2.1 Cross-Distribution Aggregation【分布シフトの横断的集約】

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.
分布シフトの主な原因の 1 つは、中核目的とは異なるモデル内の疑似相関である（Geirhos et al, 2019）。異なる領域（または異なる分布）の学習情報を最適化目標に統合することで、モデルが真実の情報と不変の

関係を学習することを我々は期待する。以下のパラグラフにおいて、まず背景として ERM を紹介し、次に学習プロセスにおいて異なる分布の損失ランドスケープを統合することで、分布シフトに対処する方法を直接学習する方法をいくつか紹介する。

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:
経験的リスク最小化 (ERM) 物体をその特徴によって効果的に識別するモデルが開発されたシナリオを考える。最適化目標は次のように表現できる：

$$\mathrm{R}(\boldsymbol{w}) = \int \mathrm{L}\Big(y, f\left(x, \boldsymbol{w}\right)\Big) d\mathrm{P}\left(x, y\right)$$

where $\mathrm{L}(y, f(x, \boldsymbol{w}))$ denotes the loss between data labels y and model outputs $f(x, \boldsymbol{w})$, while $\mathrm{P}(x, y)$ signifies the target data distribution (Vapnik, 1991).
ここで $\mathrm{L}(y, f(x, \boldsymbol{w}))$ はデータラベル y とモデル出力 $f(x, \boldsymbol{w})$ の間の損失を表し、$\mathrm{P}(x, y)$ はターゲットデータ分布を意味する (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:
とはいえ、データセットと実世界の間にはバイアスが存在することが多く、データセットから学習された特徴は、必ずしもモデルが獲得することを意図したものではない可能性がある。ERM は、このバイアスを最適化するために統計的手法で採用されている戦略である。ERM は、実世界のターゲットデータ分布にアクセスできないことを考慮すると、データセット内の経験的データは、理想的には、この未知のターゲット分布に密接に近似すべきであるという仮定に基づいて動作する（Vapnik, 1991; Zhang et al., 2018b） この文脈において、目的関数は最適化され、以下のように再定義される：

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

where l can be different examples in one training distribution or different training distributions.
ここで、 l は一つの訓練分布の異なる例、または異なる訓練分布である。

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).
上記の目的関数を最小化することで、モデルは異なる分布における不変関係を学習することができる。ナイーブ ERM は、データがターゲット・データ分布からサンプリングされるというナイーブな仮定をする。しかし、ソース分布（またはトレーニング分布）とターゲット分布の間に大きな不一致が存在する場合、深刻な汎化の問題が依然として発生する可能性がある（Szegedy et al., 2013）

Distributionally Robust Optimization (DRO)【分布ロバスト最適化（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 Dietterich, 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:
数多くの研究が、分布シフトに対する感度は、多くの場合、偽の相関関係や、中核概念とは無関係なショートカット特徴に依存することから生じるとしている（Geirhos et al, 2019; Hendrycks and Dietterich, 2018）。例えば、モデルは画像内の正しい特徴を採用するのではなく、背景の特徴に基づいて判断することがある（Geirhos et al, 2019; Beery et al, 2018）。先行研究（Ben-Tal et al., 2009; Peters et al., 2015; Krueger et al., 2021）で築かれた土台を基に、OOD 汎化は以下のように定式化できる：

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

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)

この最適化は、リスク関数集合 $\{r_e | e \in \mathcal{D}\}$ の中で最大値を小さくすることで、$\mathcal{D}$ と表記される摂動（perturbation；微小なノイズ）集合全体で最悪ケースの性能を向上させようとするものである。分布ロバスト性最適化 (DRO)（Duchi et al., 2021）では、摂動集合は異なるドメインの学習分布の混合をカバーし、上記の目的関数を最小化することで、モデルは異なる訓練分布間の不変関係を見つけることができると期待される。しかし、過剰パラメータ化されたニューラルネットワークに素朴に DRO を適用すると、最適な結果が得られない可能性があることに注意すべきである（Sagawa et al., 2020）したがって、DRO を l_2 ペナルティ（Cortes et al., 2009）や早期停止（Prechelt, 2002）のような正則化手法と組み合わせることで、汎化性能を大幅に向上させることができる。DRO の詳細については、例えば Rahimian and Mehrotra (2019); Sagawa et al. (2020); Lin et al. (2022a) を参照のこと。

Invariant Risk Minimization (IRM)【不変リスク最小化（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).

Arjovsky et al.(2019) は、多様な訓練環境にわたって非線形で不変な原因予測因子を推定する革新的な学習パラダイムを導入し、それによって堅牢（Robust）な OOD 汎化を促進した。IRM は、様々な環境にわたって確かな性能を持つ予測モデルを訓練することを目指すと同時に、疑似特徴に依存する感受性の低減を実証する。IRM は、In-variant Causal Prediction (ICP) (Peters et al., 2015) の拡張と考えることができ、これは、間接的特徴の代わりに、各特定環境内の結果につながる直接的な原因特徴を特定するための仮説検定を含む。IRM は、さらに ICP を高次元入力データを特徴とするシナリオに拡張し、そこでは変数が明確な因果的有意性を欠く可能性がある。IRM の根底にある基本的な考え方は、低い経験的損失を達成できる多くの関数に直面したとき、すべての環境にわたって強力なパフォーマンスを示す関数を選択することで、偽の特徴ではなく因果の特徴に基づく予測子（predictor）を得る可能性が高くなるということである (Murphy, 2023)。

Risk Extrapolation (REx)【リスク外挿（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.

REx の基本形は、外挿されたドメインの摂動セット（MM-REx）に対する堅牢（Robust）な最適化であり、訓練リスクの分散（V-REx）に追加ペナルティが課される（Krueger et al., 2021）訓練リスクを減らし、訓練リスクの類似性を高めることで、REx はモデルに異なるドメイン分布における不変関係を学習させる。

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:

訓練ドメイン間の分布のばらつきを増幅することで、リスクの変化を減少させ、リスクの平等性を強制することができる。CMNIST (Arjovsky et al., 2019) を例にとると、色とラベルの間の接続を確立すること

は、ロジット（logits）とラベルを接続するよりも簡単であるにもかかわらず、色の多様性を増加させることは、この疑似相関（またはショートカット特徴）を破壊し、ロジットとラベルの間の真の不変関係を学習するモデルを支援することができる。先行研究（Vapnik, 1991; Peters et al., 2017; Krueger et al., 2021）に従い、REx は以下のように定式化できる： まず、リスク関数は以下のように定義できる：

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

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:

ここで、$L(\cdot)$ は固定損失関数を表し、異なる学習ドメインまたは環境は、$P_e(X, Y)$ 分布として定式化できる。次に、MM-REx 項は次のようにモデル化できる：

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

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:

ここで、n は異なる分布またはドメインの数を表し、$\lambda_{\min}$ はリスクの外挿範囲を支配する。V-REx 項について説明すると、次のようにモデル化できる：

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

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

ここで、$\alpha \geq 0$ は、リスク削減とリスク平等の強制のトレードオフを制御する。

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.

MM-REx 項では、$\lambda_{\min}$ はほぼ $-\infty$ に設定できる。したがって、特定のドメインの損失が大きくなり、モデルが疑似相関を学習する可能性がある。MM-REx と V-REx を最小化することで、トレーニングのリスクを低減し、トレーニングの類似性を高めることができる。MM-REx と V-REx を最小化することで、学習リスクを減らし、学習リスクの類似性を高め、モデルが不変の関係を学習することを促すことができる。さらに、REx は実験的な設定（Krueger et al., 2021）、特に因果関係の識別において大きな可能性を示しており、堅牢な汎化を達成するための説得力のあるアプローチとなっている。

3.2.2 Navigation via Mode Connectivity【モード接続によるナビゲーション】

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.

クロス分布集約に関する上記の議論に続き、本節では、前提的条件としてモード接続性を紹介する。そして、主に接続性に基づくファインチューニング（Connectivity-Based Fine-Tuning：CBFT） (Lubana et al., 2023) 法について議論し、モード接続性が、少数のパラメータを変更することによって、疑似相関の代わりに不変関係に基づく予測にモデルをナビゲートすることを説明する。

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:

モード接続性 モード接続性とは、2 つ以上の異なる局所極小値やパターンを接続する、損失関数空間内の直線的な経路を特定できる現象を指す（Garipov et al, 2018; Draxler et al., 2018）。 先行研究（Benton et al., 2021; Pittorino et al., 2022; Lubana et al., 2023）に沿って、正式な定義は以下のように定義できる：

The model's loss on a dataset $\mathcal{D}$ is represented as $\mathcal{L}(f(\mathcal{D};\theta))$, where θ denotes the optimal parameters of the model, and $f(\mathcal{D};\theta)$ signifies the model trained on dataset $\mathcal{D}$. We define θ as a minimizer of the loss on this dataset if $\mathcal{L}(f(\mathcal{D};\theta)) < \epsilon$, where ϵ is a small scalar value.
あるデータセット $\mathcal{D}$ におけるモデルの損失は $\mathcal{L}(f(\mathcal{D};\theta))$ と表され、ここで θ はモデルの最適なパラメータを表し、 $f(\mathcal{D};\theta)$ はデータセット $\mathcal{D}$ で学習されたモデルを意味する。$\mathcal{L}(f(\mathcal{D};\theta)) < \epsilon$ のとき、θ をこのデータセットにおける損失の最小化と定義する。ここで ϵ は小さなスカラー値である。

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:
θ_0 がこの経路 γ に沿って変化するとき、以下の条件が一貫して維持されるような、θ_1 から θ_2 への連続的な経路 γ が存在する場合、データセット $\mathcal{D}$ での学習によって達成された最小化子 θ_1 と θ_2 は、モード接続しているとみなされる：

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

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.
本質的に、モード接続性とは、パラメータ空間内の最小化子間の接続経路を一貫して見つけることであり、損失が非常に大きい領域に入り込むことなく、損失が小さい領域を横断することを意味する。このことは、パラメータ空間内でモデルのパラメータをファインチューニングする場合でも、モデルの性能が比較的安定したまま維持され、大幅な性能劣化が緩和されることを意味する（Garipov et al., 2018）。この概念は、より効果的な最適化アルゴリズムを設計するための基礎を築き、モデルが異なるタスク間で知識と経験を共有することを可能にし、モデルの性能と汎化能力の両方を向上させる。

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.
さらに、予測を行うために同じ属性の入力を用いる場合、2 つのモデルはメカニズム的に類似していると定義することができる。いくつかの研究では、線形結合がないことはメカニズム的な非類似性を意味することが示されており、事前学習段階で学習された疑似属性を排除するには、単純なファインチューニングでは不十分であることが示唆されている（Lubana et al., 2023; Juneja et al., 2022）。 しかし、ファインチューニングによって非線形に接続された領域に対処し、それによってモデルのメカニズムを効果的に修正して、OOD の誤汎化の問題を解決することは有望である。

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:
接続性に基づくファインチューニング（CBFT） 上述したように、最近の研究では、2 つのモデル間に線形接続性がないことは、基本的なメカニズム上の非類似性を意味することが示唆されている。Lubana et al.(2023) は、同じようなデータで学習した場合、モデルは同じような推論メカニズムを発達させる傾向が

あることを発見した。これは、画像に描かれた物体ではなく、画像の背景情報に頼って分類するような、モデルに偏りが生じる重要な理由となりうる。このモデル機構がファインチューニングの過程で調整されなければ、モデルはこのような誤った属性に依存する可能性がある。この問題を克服するために、彼らはモデルのメカニズムを変更するための有効な戦略を提案している：

$$\mathcal{L}_{\text{CBFT}} = \mathcal{L}_{\text{CE}}\Big(f\left(\mathcal{D}_{\text{NC}};\boldsymbol{\theta}\right), y\Big) + \mathcal{L}_{\text{B}} + \frac{1}{K}\mathcal{L}_{\text{I}}$$

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 $\mathcal{D}_{\text{NC}}$.

ここで、元の訓練データセットを $\mathcal{D}$ とし、疑似属性 C を含まない最小のデータセットが得られると仮定する（$\mathcal{D}_{\text{NC}}$ とする）。

Besides $\mathcal{L}_{\text{CE}}$ that denotes the cross-entropy loss between model's prediction $f\left(\mathcal{D}_{\text{NC}};\boldsymbol{\theta}\right)$ 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 $\mathcal{L}_{\text{B}}$, 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.

そのほか $\mathcal{L}_{\text{CE}}$ は、モデルの予測値 $f\left(\mathcal{D}_{\text{NC}};\boldsymbol{\theta}\right)$ と実地データ（ground truth）のラベル y との間のクロスエントロピー損失を示すが、CBFT には 2 つの主要な目的がある。(1) 第一の目的は、損失ランドスケープ内で再配置することにより、モデルの基礎となるメカニズムを修正することであり、現在の最小化装置（minimizer）との線形接続を断ち切ることである。これは、バリア損失と呼ばれる $\mathcal{L}_{\text{B}}$ を最大化することで達成される。(2) 第二の目的は、元の訓練データセットに含まれる偽の属性への依存を軽減することである。これは $\mathcal{L}_I$ を最適化することによって達成され、C を必要とせずに不変の関係を発見することができる。CBFT は、モデルの部分的なパラメータを変更するだけで、疑似特徴による目的予測から真の特徴へのメカニズム転換を可能にする。

3.3 Data Distribution Interventions【データ分布への介入】

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.

アルゴリズムによる最適化の他に、訓練データの分布を実世界の要素を含むように拡張する方法も、訓練分布とデプロイ分布の不一致を減らすことができる。このセクションでは、特に敵対的圧力とマルチエージェントダイナミクスの導入に焦点を当てる。

3.3.1 Adversarial Training【敵対的トレーニング】

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).

AI システムは敵対的ロバスト（堅牢）性（adversarial robustness）の欠如に悩まされることがある。つまり、失敗するように設計された特定の入力によって、モデルのパフォーマンスが低下するのだ（Zheng et al., 2016）。これは画像（Huang et al., 2017）やテキスト（Zou et al., 2023b; Shah et al., 2023）、画像内の意味的特徴の変化（画像のセマンティック特徴の変化）(Geirhos et al., 2019; Bhattad et al., 2019; Shamsabadi et al., 2020; Casper et al., 2022）やテキストの意味的特徴の変化（(Jia and Liang, 2017)、さらには完全にゼロから生成された事例（Song et al., 2018b; Ren et al., 2020; Ziegler et al.）でも見られる。 これらの失敗モードはレッドチーミングのセクション（§4.1.3）で取り上げている。AI モデルポリシーの堅牢性に加えて、高度な AI システムの学習を支配する報酬モデルの堅牢性も重要であることは注目に値する。勾配降下最適化プ

ロセスは、報酬モデルの抜け穴を突く可能性のある敵対者と見なされる可能性があり、報酬モデルの過剰最適化と名付けられた現象が実験的に実証されている（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.
敵対的ロバスト（堅牢）性とは、AI システムの訓練分布（訓練入力が敵対的に構成されていない）とテスト分布（事例が敵対的に構成される可能性がある）のミスマッチによって部分的に引き起こされる分布シフトの失敗のケースであると考える。敵対的訓練法（Yoo and Qi, 2021; Bai et al., 2021; Ziegler et al., 2022）は、様々な方法で訓練入力に敵対的事例を導入することにより（Bai et al., 2021）、この問題を緩和し、訓練分布を拡大し、分布の不一致を解消する。

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).
敵対的学習は敵対的攻撃に似ており、最初は画像分類の設定で始まったが（Engstrom et al., 22019a）、後に幅広い設定に拡大した。視覚モデルだけでなく、言語モデル（Wang et al., 2019a; Liu et al., 2020; Ziegler et al., 2022）、視覚言語モデル（Gan et al., 2020; Berg et al., 2022）等にも敵対的学習アルゴリズムが提案されている。モデルの種類としては、分類モデル (Bai et al., 2021)、生成モデル (Ziegler et al., 2022)、RL エージェント (Pinto et al., 2017; Tan et al., 2020) に敵対的学習が適用されている。

There are two major types of adversarial training: *perturbation-based* and *unrestricted*.
敵対的トレーニングには大きく分けて摂動（perturbation）ベースと無制限の 2 種類がある。

- **Perturbation-based Adversarial Training**. Mirroring *perturbation-based adversarial attack* (see §4.1.3), perturbation-based adversarial training introduces adversarially perturbated 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.

- **摂動に基づく敵対的訓練**。摂動に基づく敵対的攻撃（4.1.3 節参照）を反映して、摂動に基づく敵対的訓練は、敵対的に摂動された事例（すなわち、モデル性能を低下させるように設計された、通常のデータ入力に対する小さな変化）を訓練に導入する（Goodfellow et al., 2014）。この系統の技術（Bai et al., 2021）には、勾配ベースの摂動入力に対するモデル性能を評価するために損失関数に正規化の条件（regularization term）を追加するベースラインアプローチ（Goodfellow et al., 2014）、教師なし（Carmon et al., 2019）または自己教師あり（Hendrycks et al., 2019）アプローチ、および訓練中に敵対的圧力を徐々に強めるカリキュラム学習の導入などの様々な補足技術が含まれる。

- **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.

- **無制限敵対的トレーニング**。無制限敵対的攻撃（4.1.3 節参照）を模倣した無制限敵対的訓練は、摂動に基づく敵対的訓練を汎化し、モデルを欺くことができるあらゆる敵対的事例を含める。これには、生

成モデルを用いて任意の敵対的入力をゼロから生成する生成的敵対的訓練（Poursaeed et al.、2021）や、構文的または意味的に修正された敵対的事例を訓練入力に加えることで、モデルの非敵対的性能への悪影響を驚くほど排除する方法（Ziegler et al.、2022; Mao et al.、2022）が含まれる。無制限敵対的攻撃に関するほとんどの研究は、無制限敵対的訓練にも適用され（概要については 4.1.3 節を参照）、無制限敵対的訓練手法の重要な部分を形成する。

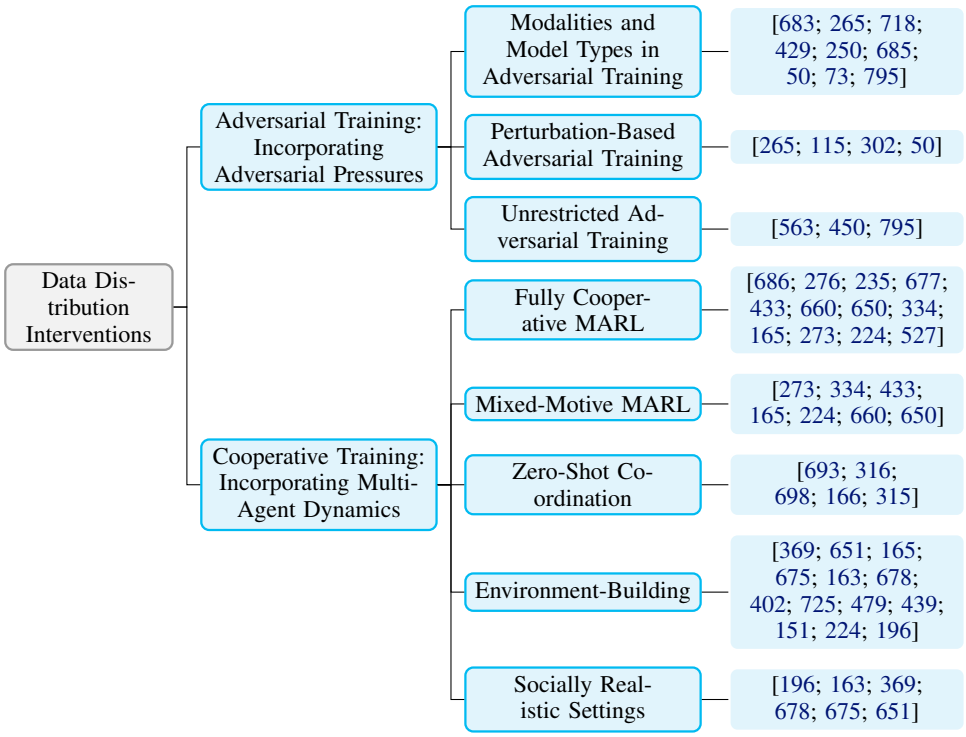

Figure 8: A tree diagram summarizing the key concepts and literature related to Data Distribution Interventions. The root node represents Data Distribution Interventions that try to combine multiple distributions during training, for example, adversarial examples and multi-agent interaction. The main 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.
図８：データ分布介入に関する主要な概念と文献をまとめたツリー図。ルートノードは、例えば敵対的事例やマルチエージェント相互作用など、訓練中に複数の分布を組み合わせようとするデータ分布介入を表す。メインブランチは、敵対的圧力を取り入れた敵対的トレーニングやマルチエージェントダイナミクスを取り入れた協調的トレーニングといった有望な手法を表している。また、協調的手法には、環境構築、社会的に現実的な設定、ゼロショット協調、その他のマルチエージェント強化学習（MARL）ベースの手法が含まれる。

3.3.2 Cooperative Training【協調的トレーニング】

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*.

協調的 AI（Dafoe et al., 2020, 2021）は、AI システムの非協力的で集団的に有害な行動に対処することを目的としている（§ 1.1.2 参照）。AI システムにおける協調能力の欠如は、分布シフトの下での失敗の一形態と見なすことができる。システムは、大規模なマルチエージェントである可能性のある現実世界とは質的に異なる単一エージェント設定で訓練される。環境に他のエージェントが存在すると、環境の状態遷移のダ

イナミクスが質的に変化し、観測と報酬の共同分布が変化するため、この違いはまさにデータ分布の違いである。我々は、協調訓練によってマルチエージェント間の相互作用を含むように訓練分布を拡張することによって、この問題にアプローチする。

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.

4.3.1 節で、マルチエージェント強化学習（Multi-Agent Reinforcement Learning：MARL）トレーニングの具体的な形態に焦点を当て、形式的なゲーム理論アプローチを補完する、協調的 AI の一分野（我々が協調型トレーニングと呼ぶもの）を紹介する。協調トレーニングの MARL 分野では、ゲーム理論分野の焦点である協調のインセンティブ（例えば、囚人のジレンマ（Phelps and Russell, 2023）のような失敗モードの緩和）とは対照的に、AI システムの協調能力（例えば、ロボットサッカーチームの協調（Ma et al., 2022））を強調する傾向がある。ここでは、学習データ分布の拡大との関連性から、MARL 分野のみを取り上げる。

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).

MARL の分野は伝統的に、完全協調型（すべてのエージェントが同じ報酬関数を共有する）、完全競争型（基礎となる報酬がゼロサムゲームを構成する）、混合動機型（報酬インセンティブが完全協調型でも完全競争型でもない、一般的サムゲーム（general-sum games）［非ゼロサムゲーム］に対応する）の 3 つのブランチに分かれていた（Gronauer and Diepold, 2022）。このうち、完全協調型と混合動機型は、協調的 AI にとって最も関連性の高い設定であり、後者はその比較的見過ごされがちな性質のために特に強調されている。（Dafoe et al.、2020）。また、ゼロショット協調（Hu et al.,2020；Treutlein et al.,2021）、環境構築（Leibo et al.,2021）、社会的に現実的な設定（Du, 2023）など、他の研究分野もカバーしている。

- **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).

- **完全協調 MARL**　MARL の完全協調設定は、すべてのエージェントが報酬関数を共有することで特徴付けられる（Gronauer and Diepold, 2022）。この統一性により、（すべてのインセンティブが完全にアラインされるため）協力インセンティブの問題を完全に無視することができ、代わりに協調による共有目標の効果的な達成に焦点を当てることができる。一般的に採用されているアプローチ (Oroojlooy and Hajinezhad, 2023) は、純粋に独立した訓練 (Tan, 1993) という基本的な解決策から、独立した訓練を分散型コミュニケーションで補うアプローチ (Foerster et al., 2016)、そしてグローバル報酬を分解して個々のエージェントの貢献度を決定する価値の因数分解（value factorization）(Guestrin et al., 2001; Sunehag et al., 2018) まで、中心性のスペクトルにある。

- **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).

- **混合動機 MARL**　MARL の混合動機設定は、協調的インセンティブと競争的インセンティブが混在

していることが特徴であり、エージェントの報酬は同一ではないが、ゼロサムでもない（Gronauer and Diepold, 2022）。これには、チーム同士が対戦するゲーム環境（Jaderberg et al., 2019）や、交渉（Cruz et al., 2019; FAIR et al., 2022）のようなより微妙な設定が含まれる。混合動機 MARL のための技術の事例には、やはり分散型から集中型へと順番に、人間の相互作用から学ぶために IRL のような手法を用いること（Song et al., 2018a）、コミュニケーションを戦略的かつ選択的にすること（Singh et al., 2018）、批評家にグローバルな情報へのアクセスを与えることで行為者－批評者（actor-critic）手法［行為者の行動を批評者が評価する強化学習のアプローチ］を適応させること（Lowe et al., 2017）などがある。

- **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.

- **ゼロショット協調**　ゼロショット協調とは、AI システムが他のエージェント（人間のエージェントを含む）と効率的に協調できるようにすることである（Hu et al., 2020; Treutlein et al., 2021）。－－見ず知らずの人間同士でも効果的に協力することができるのだから、AI システムも同じことができるはずだ。－－初期の研究は、評価（Stone et al., 2010）、ゲーム理論的・統計的アプローチ（Albrecht and Ramamoorthy, 2013）、人間のモデリング (Krafft et al., 2016) など、アドホックな調整（ad hoc coordination）という名称で発表された。最近の進歩としては、堅牢性を達成するために訓練パートナのポリシーの特定の側面をランダム化するアザープレイ（other-play） (Hu et al., 2020)、マルチレベルの再帰的推論の導入 (Cui et al., 2021)、パートナーの過去の行動を非協調的なポリシーによって取られたと解釈することによってセルフプレイの恣意的な慣習を排除する off-belief 学習 (Hu et al., 2021) などがある。

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

- **環境構築**　ゲーム環境は、例えば Hanabi（Muglich et al., 2022）、外交（Cruz et al., 2019; FAIR et al., 2022）、フットボール（Ma et al., 2022）など、協力的なトレーニングのための一般的な設定となっている。より単純化すると、ゲーム理論モデル、特に古典的なマルチエージェントジレンマに基づくモデルも、環境としてよく選ばれている（Wang and Beliaev, 2021; Christoffersen et al., 2023）。また、マルチエージェント環境のフレームワークとその一式である Melting Pot(Leibo et al., 2021) は、特に協調的 AI 研究のために設計されている。環境構築プロセスの部分的自動化を目指す教師なし環境デザインの研究もある（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).

- **社会的に現実的な設定**　協調 AI 研究は、社会的に現実的な環境（Du, 2023）にもっと焦点を当てるだと提案されている。このような環境は、AI エージェントと人間エージェントの両方を含む多数のエージェントで構成されており、エージェントの構成や相互作用の形式が非常に多様である。このビジョン

[32] This is in a similar spirit to *domain randomization* (Tobin et al., 2017).

（Critch and Krueger, 2020）には、より現実的でオープンエンドな環境の構築（Klügl et al., 2005; Lehman et al., 2008; Wang et al., 2019b; Suo et al., 2021）、MARL のスケールアップ（Sun et al., 2020; Du, 2023）、社会制度や規範などの新しい制御手段の組み込み（Singh, 2014）などが含まれるが、これらに限定されない。

4 Assurance【アシュアランス：保証】

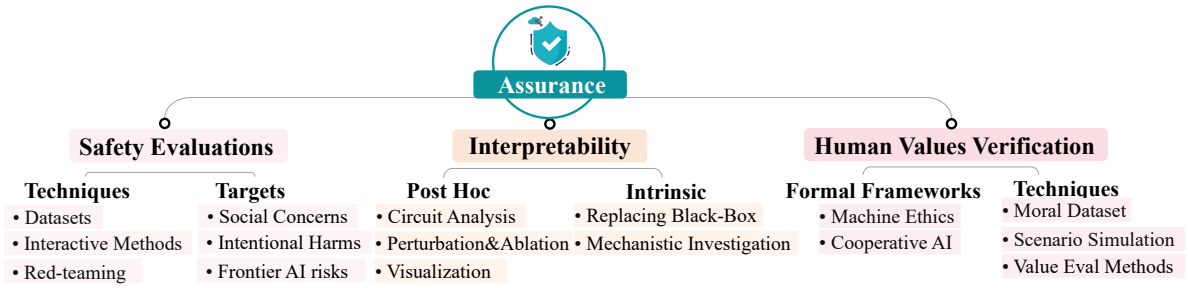

Figure 9: Our organization of research directions, techniques, and applications in assurance. We divide this section into *three* parts: Safety Evaluations – evaluation of AI systems' safety, which refers to the mitigation of accidents and harmful events caused by the AI system; Interpretability – making 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.
図９：アシュアランスにおける研究の方向性、技術、応用について整理した。このセクションは 3 つのパートに分かれている；安全性評価–AI システムの安全性を評価することで、AI システムによって引き起こされる事故や有害事象を軽減すること； 解釈可能性–AI システムやその判断プロセスを人間にとってより理解しやすくすること； 人間的価値の検証–AI システムが社会的・道徳的規範を遵守できるかどうかを検証すること。図には、これらのセクションの複雑なロジックも示されている。

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.

アシュアランスとは、AI システムが実際に訓練後やデプロイ後に、AI システムの実用的なアラインメントを測定し強化することを指す（Batarseh et al., 2021）。ここでは、一定のロジックに基づき、アシュアランスを 3 つに分類する： 安全性評価-タスク実行中の事故を最小化することをアシュアランスの基本ニーズとして AI システムを評価する、解釈可能性-AI システムの意思決定プロセスを人間が理解できることをアシュアランスすることで、評価の先にある安全性と相互運用性をアシュアランスする、人間的価値観の検証-AI システムが人間的価値観、倫理観、社会規範に適合できるかどうかを検証し、図 9 に示すような AI システムの人間社会への統合という高いレベルのニーズを満たす。AI システムが安全かどうか、アラインされているかどうかを判断する手法の他に、AI システムやその導入プロセスに積極的に介入し、その特性を保証する手法もある。

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.

機械学習解除（Machine Unlearning） モデル事前学習用のデータセットには、生物兵器やサイバー攻撃に関する情報など、様々な種類の好ましくない潜在的に危険なコンテンツが含まれている（Hendrycks et

al., 2021b）。機械学習解除の分野は、モデルの学習後にそのような知識を除去することを目的としている（Bourtoule et al., 2021） 学習データセットの直接的なフィルタリングと比較すると、このアプローチはより技術的な課題に直面するが、より柔軟な展開が可能であり、また与えられた情報の断片を分類的に除去することができる（Eldan and Russinovich, 2023）。データセットのフィルタリングと学習解除（unlearning）は、共に最適に機能する補完的なアプローチとして捉えるべきである。

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.

アラインされていないシステムのコントロール 完全なアラインメントは難しいかもしれないが、ミスアラインメントの程度が限定的で、監督する（supervisor）AI システムにアクセスできるのであれば、アラインメントされていないモデルを安全に利用することは可能である。能力の異なる信頼できるシステムと信頼できないシステムが与えられたときに、失敗の確率を最小化するためのアルゴリズム手順が開発されている（Greenblatt et al., 2023）。一般に、デプロイ手続きのアラインメントに焦点を当てたプロセスエンジニアリングは、探求すべき価値ある方向性であろう。

We then go on two review the three categories of alignment measurement efforts.
続いて、アラインメント測定の取り組みを 3 つのカテゴリーに分類してレビューする。

4.1 Safety Evaluations【安全性評価】

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 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).

安全性とは、AI システムの設計上の欠陥によって引き起こされる事故を軽減し、AI システムの意図された設計目的から逸脱する有害な事象を防止することを指す（Amodei et al., 2016）。実際、安全性はすべての工学領域に共通する要件である（Verma et al., 2010）。さらに、AI システムの特性上、AI システムの構成において特に重要である（Steinhardt, 2015）。我々は、AI システムの安全性を以下のカテゴリーに分類する： 社会的懸念（Social Concerns）とは、毒性などの側面を含む、安全な AI システムの明示的かつ比較的識別可能な特徴を指す（Stahl and Leach, 2023）。意図的行動（Intentional Behaviors）とは、権力追求、欺瞞、その他のフロンティア AI リスクに代表される、比較的複雑な調査と実質的な潜在的危害の特徴を共有する（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.

以上のロジックに従い、§ 4.1.1 で安全性評価のデータセットとベンチマークを形成する技法から始め、§ 4.1.2 でさらに評価対象とその特徴を探る。本節の最後には、評価以外の AI システムの堅牢性を評価するレッドチーミング技法 § 4.1.3 を含める。

4.1.1 Datasets and Benchmarks【データセットとベンチマーク】

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.

安全性評価に関する議論では、データセットとベンチマークを基礎的な要素として優先させることが極めて重要であるため、まずデータセットとベンチマークを構築するための基本的な手法を紹介し、その後、新しい双方向手法を紹介する。

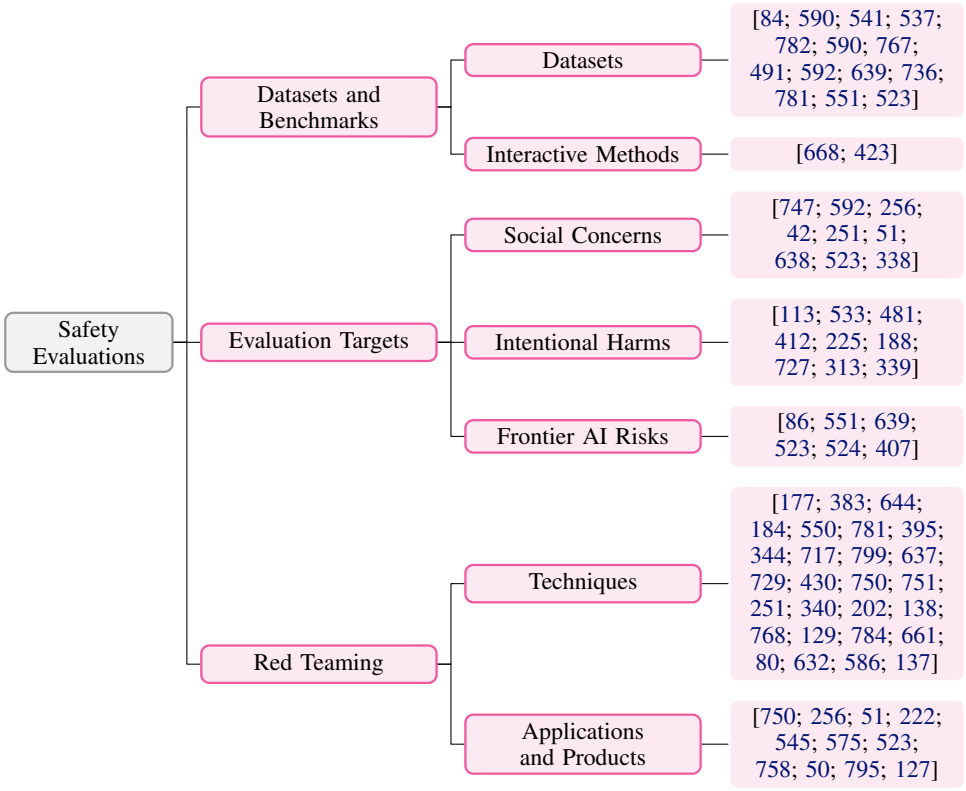

Figure 10: A Tree diagram summarizing the key concepts, logic, and literature related to Safety Evaluation. The root of the tree represents Safety Evaluation, which aims to *measure the accidents caused by design flaws in AI systems and harmful events that deviate from the intended design purpose of 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.
図 10：安全性評価に関する主要な概念、論理、文献をまとめたツリー図。ツリーのルートは安全性評価を表しており、AI システムの設計上の欠陥や、AI システムの設計目的から逸脱した有害な事象によって引き起こされる事故を測定することを目的とする。メインブランチは、データセットとベンチマーク、評価対象、レッドチーム技術など、安全性評価の主な構造を表している。さらにサブブランチは、それぞれのブランチを探求する主要な研究をリストアップしている。この図は、AI システムの安全性アラインメント度を測定するための研究の方向性と具体的な手法の概要を示している。

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.

データセット すべてのアシュアランス手法の中で、データセット法は最も初歩的で簡単なものと考えられる（Celikyilmaz et al., 2020） この手法では、あらかじめ定義されたコンテキストとタスク（Paullada et al., 2021）を AI システムに提示することで、データのコスト、質、量のバランスをとりながら、AI システムの応答を評価する。データセット法の研究には、データソース、アノテーションアプローチ、評価指標が含まれる。評価指標は対象によって異なるため（Sai et al.、2022）、本セクションでは主にデータセットソースとアノテーションの手法に重点を置く。

- **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.

- **エキスパート設計** ドメインの初期段階では、エキスパート設計がデータセットの構築に広く用いられており、エキスパートが実際のニーズに基づいてサンプルを作成し、データセットが潜在的に危険な状況を幅広くカバーしていることを確認してデータセットを形成している（Roh et al., 2019）。例えば、初期段階のデータセット、例えば偏り検出のための WEAT（Bolukbasi et al., 2016）や BBQ（Parrish et al., 2022）は、エキスパートデザインを用いて、コストと広さ（breadth）の点で制約を共有しながら、広いカバレッジと高い精度を獲得し、後のより効率的な手法の開発につながった。

- **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.

- **インターネット収集** これまでのエキスパート設計手法には、コストが高く、効率が低いという欠点がある。一方、インターネット収集は、実際のテキスト生成シナリオを反映し、実際のユーザーが生成したテキストコンテンツを大規模に取得することができる（これにより、訓練とテストの両方に便利）(Yuen et al., 2011)。しかし収集された生データは慎重な選別とアノテーションが必要である（Roh et al.、2019）。これらのデータセットのよく知られた例として、OLID（Offensive Language Identification Dataset）（Zampieri et al., 2019）と SOLID（Semi-Supervised Offensive Language Identification Dataset）（Rosenthal et al., 2021）は毒性評価のためにオリジナルの Twitter テキストを収集し、WinoBias（Zhao et al., 2018）と CrowS-Pairs（Nangia et al., 2020）はさらなるアノテーションのためにインターネットからバイアスを含む可能性のあるコンテンツを収集する。しかし、Papernot et al. (2016) でも言及されているように、インターネットから収集されたデータセットには当然、プライバシーや安全性の懸念などのリスクが伴うため、追加の処理が必要であることを認識することが重要である。

- **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.

- **AI の生成** データセットを自律的に生成するというコンセプトは、LLM の初歩的な形態が出現する以前から、比較的早い時期に検討されていた（Weston et al., 2015） しかし、この初期の段階では、AI が生成したデータセットは AI システムの能力によって制限されていたため、その品質はインターネット上で収集され、手作業でアノテーション（注釈）が付けられたデータセットほど良くなかった。LLM が既存のデータセットの構造や論理を模倣して新しいデータセットを構成する能力を獲得したのは、LLM が論理的推論の文脈理解において比較的高いレベルに到達し、人間レベルの性能に近づいたり凌駕したりするようになってからである（OpenAI, 2023a）。Zhang et al. (2022) や Perez et al. (2023) などの論文で示されているように、AI システムは評価目的のデータセット生成において進歩を遂げ、いくつかの古典的なデータセットの品質を凌駕している。しかし、これらの論文によれば、このアプローチは、命令の誤解や事例の多様性といった問題を含め、大規模モデル自体の能力に根ざした限界に依然として直面しており、さらなる改良が必要である。

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.

インタラクティブな手法 データセットは静的な性質を持っているため、評価内容が比較的固定されており、的を絞ったトレーニングに対して脆弱である可能性がある（Holtzman et al., 2019） さらに、評価内容は、対応する能力の長所と短所を完全に反映していない可能性がある（Engstrom et al., 2020）。言語モデル評価に対する要求がエスカレートし続ける中、新しい双方向のアシュアランス手法が登場し、それらは 2 つのグループに分類できる：「監督者としてのエージェント」と「環境との相互作用」である。

- **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.
- **監督者としてのエージェント** これは、AI モデルの出力を評価するためにエージェントを使用するアシュアランス手法である。この評価手法は、そのダイナミズムと柔軟性が特徴である。一般的に、エージェントと評価対象の AI システムとの間のインタラクションには、あらかじめ定義されたフレームワークが存在する（Cabrera et al., 2023）。この方法では、エージェントは、オンラインシステムを通じて実験に従事する人間の参加者であったり（Stennon et al., 2020）、マルチターン（multi-turn）相互作用を通じて比較的能力の低い言語モデルを評価するより高度な言語モデルであったり（Lin and Chen, 2023）、スケーラブルな監視の文脈では、能力は低いがより信頼できるモデルであったりする（Greenblatt et al., 2023）。この評価形態は、人間のエージェントと比較して、自動化や低コストといった利点を提供する。

- **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.
- **環境との相互作用** これは、人間や他の LLM などの要素を用いて比較的現実的な環境を作り出し、複数回の相互作用を通じて AI モデルのアラインメント品質を評価することを目的としている（Liu et al., 2024b）。その一つの方法として、複数の LLM が対話を行うピアディスカッションを用いることで、特にお互いの能力が比較的近い場合に、AI システムの評価を高めることができる。さらに、ワールドモデルを構築することにより（Li et al., 2022b）、AI システムの汎化能力や探索能力を総合的に評価することができる。

4.1.2 Evaluation Targets【評価対象】

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.

安全なアラインメントという目標を達成するために、AI システムのアシュアランスは様々な小さな目標に分けることができる（Shevlane et al., 2023）。以下のセクションでは、これらのテーマについて紹介し、さらに、これらの領域におけるアシュアランス手法のドメイン固有の分析について議論する。表 3 は、これらの領域におけるアラインメントアシュアランスの事例を示す。

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)

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.
表３：安全性評価事例のチャート： 具体的なデータセットの作品を、その詳細情報とともに一覧にしている：評価対象、初公開時期、最新更新時期（データセットによってはコンスタントに更新されるため、分けて記載）、情報量（情報形式単位の合計）、機関、情報形式、ベースラインモデル、情報源。また、より多くの情報を記載するため、表中の略称を一部変更した： また、より多くの情報を記載するために、リリース時期や最近の更新は、年月の下２桁を連結して短くし、論文の筆頭著者の所属機関のみを取り、情報形式の長い単語は大文字の組み合わせで置き換えている： SP はセンテンス・ペア、SL はセンテンス・ラベル、ST はセンテンス・テンプレート、PP は代名詞（pronoun）ペア、SS はシングル・セレクションである。

	Dataset	Release Time	Recent Update	Info Quantity	Institution	Information Form	Baseline Model	Infomation Source
Bias	Aequitas [607]	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 [728]	18/05	-	8.9K	Google	PP	Transformer	Wikipedia
	OLID [767]	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 [747]	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 [610]	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 [727]	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

employ a red teaming design pattern that induces toxic responses through adversarial inputs, further strengthening the assurance of model robustness.

毒性 AI システムの出力に含まれる、人間にとって役に立たない、あるいは有害な内容を指す（Sheth et al., 2022）。 高度な言語モデルが登場する以前、初期の毒性評価は、主に有毒な言語を検出し、インターネットのコンテキストで有害なステートメントを識別することに重点を置いていた。それには WCC（Wulczyn et al., 2017） があり、Wikipedia のディスカッションページからコメントを収集し、手作業でラベル付けした。事前学習済み言語モデルの出現により、毒性に対するアシュアランスは、言語モデルが特定のプロンプトに応答して有害なコンテンツを生成するリスクを評価するプロンプト生成パラダイムを採用した（Gehman et al., 2020; Ganguli et al., 2022; OpenAI, 2023a）。しかし、クラウドソーシング環境では、アノテーションのスコアが人によって異なる可能性があるため、クラウドソーシングの質を高めるには、クラウドソーシングワーカーがチャット中に２つの異なる回答から選択する、リレーショナルラベリングが必要である（Bai et al., 2022a） さらに、その後のデータセット（Ganguli et al., 2022; Ji et al., 2024b）は、敵対的入力によって有害な回答を誘導するレッドチーミングデザインパターンを採用しており、モデルの堅牢性のアシュアランスをさらに強化している。

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.

権力追求 AI システムが一定レベルの知能を持つようになると、人間に対する権力を求めるようになる可能性があることは、一種のリスクである（Turner et al., 2021）。 Carlsmith（2022）で著者らは、AI システムは、高度な能力、エージェント的計画、戦略的認識など、権力追求のための条件をすでに備えていると指摘している。しかし、権力追求に対抗するアシュアランスはまだ初期段階にある。この分野の代表的な研究として、マキャベリ（Machiavelli）（Pan et al., 2023a）がある。これは、意思決定ゲームからなるベンチマークを構築し、AI システムがゲーム中に競争と道徳倫理のバランスを取ることができるかどうかを評価するものである。この研究の結論は、AI システムは報酬の獲得と道徳的な行動のバランスをとることにまだ苦労していることを示唆しており、したがってこの分野でのさらなる研究が必要である。

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).

欺瞞的アラインメント AI システムが状況を認識している場合、高い報酬を得ることで大幅な勾配降下を防ぎ、本来の目標を維持することができると認識することがある（Hubinger et al., 2019c; Kenton et al., 2021; Ngo et al., 2024）。 このプロセスは欺瞞的アラインメント（Deceptive Alignment）と呼ばれる。現在の文脈では、Hubinger et al.(2024) によって証明されているように、欺瞞的アラインメントはすでに達成可能である。欺瞞的アラインメントを直接評価することは困難であり、代名詞（the pronoun）の欺瞞的アラインメントは伝統的な訓練と評価のループに反するからである。従って、欺瞞的アラインメントは、モデルパラメータの解釈（4.2 参照）や表現工学（representation engineering）（Zou et al., 2023a）のような間接的な方法によって発見されるかもしれない。

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.

さらに、欺瞞的アラインメントは状況認識（situational awareness）と密接に関連している。つまり、AI システムは、対応する決定を行うために、作業環境内のエンティティの状態や発展についてある程度の予測と理解を持つ。Li et al.（2022b）では、著者らはボードゲーム・オセロにおける言語モデルの性能を評価し、言語モデルが非線形表現において行動空間内の将来起こりうる状態を予測する能力を持つことを示している。

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).

ハルシネーション（幻覚） AI システムは、事実の知識やデータに基づかない情報や応答を生成することがあり、誤解を招いたり、誤った内容を生成したりすることがある（Ji et al., 2023）。ハルシネーションの評価は、AI システムの出力に含まれる知識と、その学習データや知識ベースによって与えられる知識との整合性を保証することを目的としている (Ji et al., 2023; Zhang et al., 2023c)。統計に基づく最も初期のハ

ルシネーション評価手法は、n-gram［テキストドキュメント内の連続する n 個のアイテムの集合］を使用して、入力内容と出力内容の語彙の重なりを直接計算していた（Dhingra et al., 2019; Wang et al., 2020）。しかし、このタイプの評価には限界がある： 語彙の重なりのみを考慮し、意味や文意を考慮しないのである（Ji et al., 2023）。より複雑なハルシネーションの評価には不向きである。その後のアシュアランス手法は、統計的アプローチからモデルベースの方法へと移行し、統計的トークン差に基づく方法と比較してより堅牢になった（Honovich et al., 2021） この評価法は以前のものよりも進歩しているが、モデルが出力できるのはハルシネーションの程度のみであり、特定のエラーをピンポイントで特定するのは難しいという限界が残っている（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:

フロンティア AI リスク 上記のアシュアランス内容に加え、近年の AI システムの高度化により、新たなアシュアランスニーズが次々と生まれている（OpenAI, 2023a）。現在のところ、これらのアシュアランスの必要性に関する研究に利用可能な公開情報は多くないため、本セクションでは、より重要なものについて簡単に紹介する：

- **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.

- **サイバーセキュリティと生物兵器** 高度な LLM は、サイバー攻撃や生物兵器の製造、その他の極めて有害な行動に悪用される可能性がある（Shevlane et al., 2023）。GPT-4 はコンテキストウィンドウ（context window）が限られているため、ネットワークの脆弱性を悪用する上で重要な 役割を果たすことはできないが、ネットワークの脆弱性を特定したり、ソーシャルエンジニアリングにおいて 強力な能力を発揮することが証明されている（OpenAI, 2023a）。同様に、Lentzos (2022) は、生物兵器や軍事分野における AI システムの強力な能力を述べ、そのような能力が悪用されるリスクを強調している。また、これらのモデルが悪意のある要求を識別し、拒否できるようにする必要性を強調している。

- **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.

- **欺瞞と操作** AI システムは、虚偽の情報を流布したり、人間を同調させたり、人々の信念や政治的影響を形成したりするなど、テキストを出力することでユーザーに悪影響を及ぼす可能性がある（Shevlane et al., 2023; Sharma et al., 2024） 。ハルシネーションとは異なり、ここでの誤情報はモデル自体の欠陥ではなく、むしろ意図的な行動である。この種の行動を制御するために、特別なアシュアランス措置を設計する必要がある。

- **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.

- **脱獄（Jailbreak）** 脱獄とは、例えば特定の種類の入力を構成することで、ユーザーによって AI システムのセーフガード・メカニズムが迂回されることを指す。この行動は、テキストに限定されることもあれば（OpenAI, 2023a; Deng et al., 2023; Huang et al., 2024; Yong et al., 2023）、マルチモーダルな形態をとることもある（OpenAI, 2023b）。具体的には、マルチモーダルな脱獄は、攻撃コンテンツを識別するための伝統的なテキストベースのヒューリスティックな方法を実行不可能にし、特別なマルチモーダル処理方法を必要とする。脱獄のさらなる議論は、§ 4.1.3 で見つけることができる。

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

- **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]

- **自己保存と増殖**　これは、AI システムが自己防衛と複製を求める傾向のことであり、この過程で環境からの限界を突破する。これらの傾向は、手段的副目標の事例である（Bostrom, 2012）。この傾向は有益に活用できる一方で、規制がない場合には危険である（Perez et al., 2023） この傾向は様々な情報源によって強調され、評価されてきた（Perez et al, 2023; Kinniment et al, 2023; OpenAI, 2023a,b）。

4.1.3 Red Teaming【レッド・チーミング】

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).

レッド・チーミングとは、AI システムがアラインされていない出力や行動（例えば、欺瞞や権力追求などの危険な行動、その他に、有害な出力や偏った出力などの問題）を出すように誘導するシナリオを生成し、これらのシナリオでシステムをテストする行為である。その目的は、敵対的な圧力をかけることによって、つまり具体的にシステムを失敗させようとすることによって、システムのアラインメントの堅牢性を評価することである。一般的に、言語モデルや視覚モデルを含む最先端のシステムは、このテストに合格しない（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.

ゲーム理論などの分野では、レッド・チーミングはより早くから導入されており、計算機科学の分野でも、システムの信頼性や堅牢性を敵対的に評価するという同様の意味を持つ、セキュリティ分野でレッド・チーミングの概念が提唱されていた。その後、Ganguli et al. (2022) と Perez et al. (2022) が、この考え方を AI の分野、より具体的にはアラインメントの分野に導入した。

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.

レッド・チーミングの動機は 2 つある。(1) 訓練されたシステムのアラインメントをアシュアランスするため、(2) 敵対的な訓練（Yoo and Qi, 2021; Bai et al., 2021; Ziegler et al., 2022）、プロービング・モデル（probing models）（Kalin et al., 2020)）、およびさらなるユーティリティのための敵対的な入力源を提供するためである。 ここでは前者に注目する。この 2 つの目的は分離可能ではないことは注目に値する。第一の動機をターゲットとした研究は、第二の動機の基礎を提供する助けにもなる。

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.

強化、最適化、誘導、逆の文脈生成（Reinforced, Optimized, Guided, or Reverse Context Generation）
このカテゴリには、言語モデルからミスアラインメントを誘導する首尾一貫したコンテキスト （プロンプト）を生成するために、さまざまな方法を使用することが含まれる。Perez et al. (2022); Deng et al. (2022);

Casper et al. (2023c) は、RL を使用して別の言語モデルを訓練またはチューニングし、望ましいプロンプトを生成させる。Perez et al. (2022); Si et al. (2022) は、ゼロショット、少数ショット、教師ありファインチューニングベースの生成など、他の方法も使用している。Lee et al. (2022) と Jones et al. (2023) は、プロンプトに対して最適化（それぞれベイズ 最適化と離散最適化）を実行することで、ミスアラインメントを誘発するコンテキストを生成する。Dathathri et al. (2019) と Krause et al. (2021) は、より小さな分類器を用いて LLM の 生成を誘導する方法を提唱している。これは無害化の手法として提案されているが、レッドチーミングのコンテキストにも転用可能である。最後に、Zhang et al. (2022) は、逆生成、すなわち、与えられた応答を条件として敵対的コンテクストを構築することで、ミスアラインメントを誘発するコンテクストを生成する。

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.

手動脱獄と自動脱獄 上記の 4.1.2 で定義されているように、脱獄（Shen et al., 2023）とはインフォーマルな用語であり、製品のユーザーに対する制約を回避する行為、LLM の場合は、アラインメント訓練の成果であるミスアラインメントを誘発する質問に答えないという LLM の傾向を回避する行為を指す。既存の試みのほとんどは、非公式な報告の形でインターネット上に散在しており、原文に接頭辞や接尾辞を追加するものを含む (Zou et al., 2023b)。研究では、既存の試みを記述的に分析し（Liu et al., 2023; Shen et al., 2023; Deng et al., 2023; Huang et al., 2024）、現象の因果関係を説明している（Wei et al., 2024）。さらに、過去 (Wallace et al., 2019) や現在 (Zou et al., 2023b; Shah et al., 2023) の研究では、LLM がミスアラインメントを誘発する質問を避ける傾向を無効化するようなプロンプト、接頭辞、接尾辞を自動的に生成する効果的な方法が提案されている。

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.

クラウドソーシングによる敵対的入力 いくつかの研究（Xu et al., 2020, 2021; Ganguli et al., 2022）では、クラウドソーシング、つまり人間のレッドチーム参加者を（おそらくオンラインプラットフォームを通じて）募集し、彼らに敵対的なプロンプトを提供するよう指示することで、ミスアラインメントを誘発するプロンプトを生成している。また、AI 業界の企業も、OpenAI のレッドチーミングネットワークや Google のバグハンタープログラムなど、敵対的なインプットを収集する仕組みを構築している。これらの方法は（間違いなく）より柔軟性があり、実際のユースケースに似ているが、コストが高く、スケーラビリティが低い。

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).

摂動に基づく敵対的攻撃 コンピュータビジョンの分野では、摂動、すなわち画像のピクセルコンテキストに小さな摂動（通常はピクセル単位の行列ノルムで抑制される）を与えることで、摂動された画像に対してモデルが確信を持って偽の出力を出すようにする手法（Chakraborty et al., 2021）に依拠した、ビジョンモデルに対する敵対的攻撃の研究が数多く行われてきた。この種の敵対的攻撃は、言語モデル（Jia and Liang, 2017; Ebrahimi et al., 2018; Zang et al., 2020; Cheng et al., 2020）や視覚言語モデル（Zhao et al., 2024）にも拡張されている。

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

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).

無制限敵対的攻撃 Song et al. (2018b) で提案された無制限敵対的攻撃は、より一般的な敵対的攻撃の形式である。これは、敵対事例に対するすべての制限を取り除くものである、したがって、例えば摂動に基づく手法のように既存の事例から生成するのではなく、ゼロから敵対的事例を生成することができる。無制限敵対的攻撃の手法は数多く提案されており、代表的なものとしては、生成モデルを用いて現実的な敵対的画像を生成する（Song et al., 2018b; Chen et al., 2024）や、色やテクスチャといった意味的に意味のある特徴を操作する（Bhattad et al., 2019; Shamsabadi et al., 2020）などがある。無制限敵対的攻撃は、テキスト分類モデルにも拡張されている（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).

レッド・チーミングのためのデータセット IMAGENET-A および IMAGENET-O データセット (Hendrycks et al., 2021c)、BAD データセット (Xu et al., 2020)、HH-RLHF データセットのレッドチーミングセクション (Bai et al., 2022a)、Real Toxicity Prompts データセット (Gehman et al., 2020) など、レッドチーミングや関連するトピックに関する多くの研究が、レッドチーミングのプロンプトやダイアログからなるデータセットを編集している。

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.

産業界における既存のレッド・チーミングの実践 AI 業界では、レッド・チーミングの実践が人気を集めている。採用事例としては、OpenAI（GPT-4 システムでレッドチーミングを行い、システムカードの一部を作成した）（OpenAI, 2023a）、NVIDIA（Pearce and Lucas, 2023）、Google（Fabian, 2023）、Microsoft（Ram Shankar Siva Kumar, 2023）などがある。DEF CON 31 カンファレンスのイベントでは、9 社のモデルがカンファレンスの参加者からレッドチーミングを受けた。このレッドチーミングイベントは、ホワイトハウスを含む米国公共部門の 4 つの機関と提携して開催されている。

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).

下流工程への応用（Downstream Applications） レッド・チーミングは、敵対的な入力（Yoo and Qi, 2021; Bai et al., 2021; Ziegler et al., 2022）を提供することで、AI システムの敵対的訓練において重要な役割を果たす。さらに、レッド・チーミングから生成される敵対的事例は、モデルの解釈にも使用できる (Casper et al., 2022)。

4.2 Interpretability【解釈可能性】

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

[36] https://www.airedteam.org/

[37] For a more comprehensive review of interpretability and its methods, we recommend (Räuker et al., 2023).

to sub-fields and purposes (Doshi-Velez and Kim, 2017; Rudin, 2019). There are several ways to break down interpretability research:
解釈可能性は、機械学習システムとその意思決定プロセスを人間にとって理解可能なものにする研究分野である（Doshi-Velez and Kim, 2017; Zhang and Zhu, 2018; Miller, 2019）。解釈可能性の研究は、モデルについて何か新しいことをより良く記述したり予測したりするためのツールボックスを構築する。本稿では、アラインメントと安全性に最も関連する研究に焦点を当て、経験的に、それらの技術はニューラルネットワークの内部構造と表現を研究することによってニューラルネットワークをより安全にする（Räuker et al., 2023）。ホワイトボックス・システムの安全性をアシュアランスすることは、原則的に、ブラックボックス・システムよりも容易であるため、解釈可能性は研究の重要な方向性である。解釈可能性ツールの分類法は、下位分野や目的によって異なる（Doshi-Velez and Kim, 2017; Rudin, 2019）。解釈可能性研究を分類する方法はいくつかある：

- *Explainability and Transparency*. Explainability research aims to understand why models generate specific output, whereas transparency aims to understand model internals (Critch and Krueger, 2020).
- **説明可能性と透明性**　説明可能性の研究は、モデルが特定の出力を生成する理由を理解することを目的とし、透明性の研究は、モデルの内部を理解することを目的としている（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).
- **重み、ニューロン、サブネットワーク、表現（Weights, Neurons, Sub-networks or Representations）**
この分類は、その手法が計算グラフのどの部分を説明するのに役立つかを見ることによって、解釈可能性手法を整理したものである：重み、ニューロン、サブネットワーク、潜在的表現（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).
- **安全性、あるいはディープラーニングの科学**　研究者達は AI システムを安全に展開するために解釈可能性研究を行い、またある研究者はニューラルネットワークの完全な科学を目指している。しかし、機械的な解釈可能性研究はその両方を目指すため、境界線は曖昧になる（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.
- **内在的解釈可能性と事後的解釈可能性**　解釈可能性研究は、介入の段階によって、内在的解釈可能性と事後的解釈可能性（Carvalho et al., 2019）に分けられる。前者は、内在的に解釈可能なモデルを作ることに焦点を当て、後者は、モデルの行動に対して説明を提供する事後的解釈可能性の手法を設計する。
（※ Intrinsic Interpretability を内在的解釈可能性と訳出）

- *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.
- **機械的解釈可能性、表現工学、概念に基づく解釈可能性**　AI の安全性とアラインメントのコミュニティでは、3 つの研究課題が注目を集めている（Impact, 2023）：機械的解釈可能性（Mechanistic Interpretability）は、ボトムアップのアプローチで、ニューラルネットワークによって実装されるアルゴリズムの低レベルのメカニズムを理解することを目的とする（Olah et al., 2020）、表現工学（Representation Engineering）は、対照的に、トップダウンのアプローチで、ニューラルネットワークの高レベルの認知現象を監視（操

作）する（Zou et al., 2023a）、そして概念に基づく解釈可能性（Concept-based Interpretability）は、学習された知識表現を、モデルが出力するものとは対照的に、ニューラルネットワークに位置付ける（Meng et al., 2022a,b）。この 3 つに共通するのは、特徴量とニューロンの集合を同時に結びつけることである。

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.

本節では、他の分類手法と比較して、ニューラルネットワークに限らず様々な AI システムに適した汎用的な枠組みを提供し、システムの設計中とシステムのデプロイ後に解釈可能性の分析を分けることができるため（Räuker et al., 2023）、内在的解釈可能性分類手法と事後的解釈可能性分類手法を採用した。具体的には、モデル設計段階と事後段階で行われるメカニズム的解釈可能性技法について、事後的と内在的の解釈可能性のサブセクションで分けて論じた。

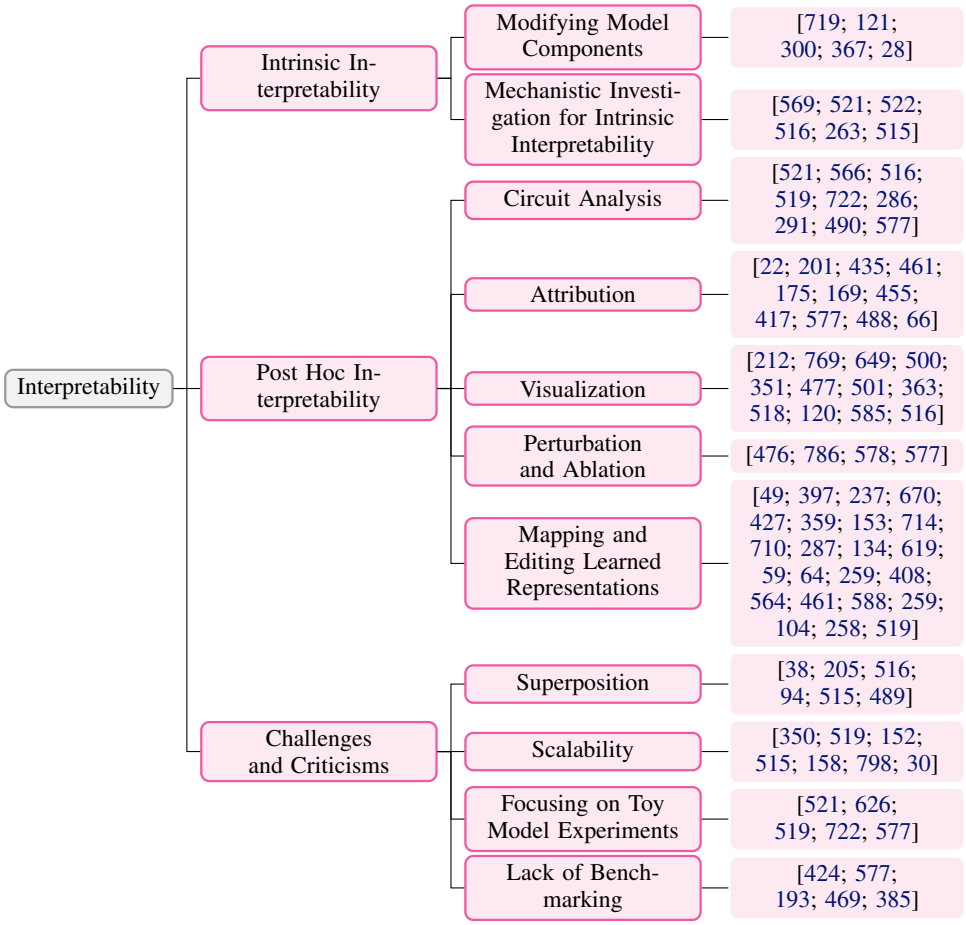

Figure 11: A Tree diagram summarizing the key techniques concepts, challenges, and literature related to Interpretability.
図 11：解釈可能性に関連する主要な技術の概念、課題、文献をまとめたツリー図。

4.2.1 Intrinsic Interpretability【内在的解釈可能性】

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.

研究者たちは、深層学習モデルをより内在的に理解できるようにする。これは通常、内在的解釈可能性（Carvalho et al., 2019 ）と呼ばれる。解釈可能なモデルの作成を重視する記号的アプローチとは対照的に、最新の深層学習アプローチでは、機能は強化されているものの、潜在的に解釈可能性が低下したモデルが得られる傾向がある。ブラックボックス化されたモデルを解釈することに比べ、内在的に解釈可能なモデルを設計することは、より安全で効率的である（Rudin, 2019）。内在的に解釈可能なモデルを作るために、研究コミュニティは、敵対的攻撃に対して堅牢で、重ね合わせ（Anthropic, 2022; Räuker et al., 2023）がないモジュラーアーキテクチャを設計する。注目すべきは、機械論的解釈可能性は、しばしば事後的解釈可能性技術のセットとみなされるが、間違いなく、より解釈可能なモデルを作るプロセスを促進する。

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).

モデル・コンポーネントを修正する フィードフォワード層（feedforward layers）などのモデル・コンポーネントは、解釈するのが難しい（つまり、その動作を人間が理解できる言葉で明確にするのが難しい）。というのも、これらの層は、無関係な入力に反応する多くの多義的なニューロンを持っているからである（Du et al., 2019）。したがって、リバースエンジニアリングを容易にするために、これらのバックボックスコンポーネント（back-box components）とその関連構造に適用される特定の変更によって、結果、それらの解釈可能性が向上する（Carvalho et al., 2019）。損失関数の修正（Ross et al., 2017）、特別な解釈可能フィルターや埋め込み空間の追加（Zhang et al., 2018c; Wang et al., 2021）、入力に応じた動的重みの使用（Foerster et al., 2017）、中間層の修正（Li et al., 2022a）などにより、解釈可能な結果を促す既存の研究が数多くある。具体的に、Lage et al.（2018）は、主観的概念を定量化するために人間のフィードバックを直接利用するヒューマンインザループ（human-in-the-loop）アルゴリズムを提案し、より信頼性の高い結果を達成している。トランスフォーマーモデルにおいて、Anthropic は活性化関数を置き換えるために SoLU を提案し、解釈可能（interpretable）ニューロンの数を増やし、性能を維持しながらリバースエンジニアリングを容易にする (Anthropic, 2022)。これは、潜在的に重要な研究分野としてまだ初期の段階であり、この手法のスケーラビリティなどの課題が残っている（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:

モデルアーキテクチャのリエンジニアリング 既存のモデルコンポーネントを修正することは、リバースエンジニアリング（Carvalho et al., 2019; Foerster et al., 2017）には有益であるが、モデルを完全に理解できるようにすることはできないため、一部の研究者は、理論的に解釈可能なモデルを構築するために、モデルアーキテクチャのリエンジニアリングを開始した（Carvalho et al, 2019; Mascharka et al, 2018）。特筆すべきは、一般的に、同じモデルの複雑さにおいて、モデルの解釈可能性とその性能の間にはトレードオフが存在すると考えられており（Alvarez Melis and Jaakkola, 2018）、これら 2 つの要素のバランスを取るモデルを設計すること、あるいはさらに、解釈可能なモデルと最先端のモデルとのギャップを縮めることが極めて重要になる（Alvarez Melis and Jaakkola, 2018; Carvalho et al., 2019; Fan et al., 2021; Espinosa Zarlenga et al., 2022）。詳細な研究の取り組みについては後述する：

- *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).

- **透明な推論ステップの作成** 推論モデルにおいて、透明なマイナーステップ（minor steps）を作成することは、モデルを解釈可能にするために極めて重要であり（Hudson and Manning, 2018）、多くの論文が、MAC（Memory, Attention, and Composition）セルを導入して記憶と制御を分離したり（Hudson and Manning, 2018）、他のアテンションベースの手法を利用したり（Lin et al., 2019; Arik and Pfister, 2021）、複雑な推論プロセスを分解したりすること（Mascharka et al., 2018）で、それを達成した。モデルの複雑性が増し、パフォーマンスにおいてある程度のコストが生じたが、これらの方法によって推論プロセスの解釈可能性を大幅に向上させ、解釈可能なモデルと最先端のモデルとの性能のギャップを縮めた。(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 (**?**)

- **複雑な知識を蒸留する** ディープ・ニューラル・ネットワークのような複雑なモデルは、しばしば高い性能を持つが、その意思決定プロセスの透明性に欠け、解釈を難しくする（Li et al., 2020）。知識蒸留は、このような複雑な「ブラックボックス」モデル（教師）から、よりシンプルで解釈しやすいモデル（生徒）に知識を伝達することで、この課題に対処する。この構造をモデル設計に導入することで、生徒モデルはより高い透明性を提供しながら教師のパフォーマンスに近似することができ、高度な機械学習モデルの能力を犠牲にすることなく解釈可能性を高めることができる（Zhang et al., 2020b; Li et al., 2020）。しかし、この解釈可能性は部分的なものであり、特に複雑なミッションにおいては、抽出された知識の解釈はまだ難しいかもしれない（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.

さらに、予測と説明の両方を提供できる自己説明モデル（Self-Explaining Models） (Elton, 2020) は、解釈可能モデル（Interpretable Models）のより良い代替方法として提案され、多くの論文がこれに取り組んでいる (Alvarez Melis and Jaakkola, 2018; Rajagopal et al., 2021)。言語モデルの場合、思考の連鎖（CoT）生成（Wei et al., 2022）は、一種の自己説明の手法として認識されるかもしれない。

4.2.2 Post Hoc Interpretability【事後的解釈可能性】

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.

このセクションでは、モデルがトレーニングされ、デプロイされた後に、モデル内部を理解するために適用される技法と手法を探る。したがって、これらの技法はしばしば事後的解釈可能性（Räuker et al., 2023）と呼ばれる。その目的は、ブラックボックス・ニューラルネットワークの低レベルの構造とユニット、およびそれらがマクロな動作と出力に及ぼす因果的な影響を理解することである。

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).

回路分析（Circuit Analysis） 回路とは、神経回路網の中で特定の機能を割り当てられるサブネットワークを指す。神経科学における対応物質である神経回路が解剖学的かつ機能的な実体であるように（Purves et al, 2001）、回路もまた物理的かつ機能的である（Olah et al, 2020）。機械的解釈可能性の研究者は、モデルの振る舞い（マクロ）を理解するために、神経回路網（ミクロ）に回路を位置づける。複数の回路が報告されている：曲線検出器のための曲線回路（OpenAI, 2021a）、文脈内学習のための誘導回路（Olsson et al, 2022）、センテンス内のオブジェクトを識別するための間接オブジェクト識別回路（Wang et al, 2022）、Python 関数のドキュメンテーション文字列（docstring）で繰り返される引数名を予測するための Python docstrings（Heimersheim and Jett, 2023）、grokking［過学習後に急激に汎化誤差が減少する（正解率が上昇する）現象］（Nanda et al, 2022）、多桁加算（multi-digit addition）（Nanda et al, 2022）、大なり記号（greater than）といった数学的能力（Hanna et al, 2024）などである。注目すべきは、現在までに行われた多くの回路分析が、簡易的な試験モデルや簡易的な試験タスク（toy models and toy tasks）に焦点を当てていることである（Räuker et al, 2023）。言語モデルの自然な動作をリバースエンジニアリングする最大の試みは、GPT-2 Small にあり、28 個のヘッドを持つ間接物体識別回路を見つけたことである（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).

プロービング（Probing） プロービングとは、興味ある内部学習された表現に対して独立した分類器を訓練し、概念／特徴を抽出する技術の集まりである。一例として、Gurnee らは隠れ層における空間と時間の線形表現を研究するためにプロービングを用いた (Gurnee and Tegmark, 2023)。プロービングは隠れ層を理解するために研究者に好まれているが（Alain and Bengio, 2017）、限界がある。一つは、プロービングは隠れ層の学習された表現を理解するのに役立つが、学習された表現が予測値を生成するためにモデルによって使われているかどうかはわからない（Ravichander et al., 2021; Belinkov, 2022）。もう一つは、データセットの問題がモデルの問題と混同される可能性がある（Belinkov, 2022）。安全性とアラインメントの文脈では、プローブを訓練するには、データセットに関心のある概念／特徴が含まれている必要がある。つまり、分布外の特徴（モデルによって学習されたと思われるが、そのデータセットがないという特徴）を検出するためにプローブを使用することはできない。注目すべきは、プロービングの文献に基づいて構築された表現工学が、道徳、感情、嘘、権力追求行動など、高レベルの認知現象や危険な能力を検出するために導入されていることである（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).

辞書学習（Dictionary learning） 事後的解釈可能性の重要な課題は、重ね合わせ、すなわち、ニューロンは人間が解釈可能な複数の特徴を同時に符号化する傾向があるため、個々の特徴を識別することが非常に困難になることである（Elhage et al., 2022）。この課題に対処するため、教師なしかつスケーラブルな方法でこれらの特徴を分離する、辞書学習に基づく方法が提案されている（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).

モデル帰属（Model Attribution） 帰属は、ニューロン応答とモデル出力に対する、いくつかの構成要素（ヘッド、ニューロン、層、入力を含む）の寄与を調べる一連の技法である（Räuker et al., 2023）。勾配に基づく帰属は、解釈の質を評価し、モデルが学習した事実を探るために導入された（Ancona et al., 2018; Durrani et al., 2020; Lundstrom et al., 2022; Dai et al., 2022）。しかし、これらの方法は因果関係を説明できないため（Räuker et al., 2023）、限界がある。直接的ロジット（logits）帰属は、次のニューロン（Lieberum et al., 2023; McGrath et al., 2023; Belrose et al., 2023; Dar et al., 2023）の予測に対する個々のニューロンの直接的な寄与を識別することである。しかし、帰属手法もまた、顕著な制約に悩まされる。それは、興味のある特徴のデータセットがあるシナリオにしか役立たないということである。その結果、このような帰属手法は、分布外（OOD）特徴（いくつかの誤判定シナリオを含む）の理解には役立たない (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.

データの帰属 ある行動につながる訓練データの部分集合を特定することで、当該行動の安全性と、その行動を奨励または防止する方法の両方に関する知見を得ることができる。このような帰属を行う方法として、影響関数を用いた方法 (Koh and Liang, 2017; Grosse et al., 2023) が提案されている。

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).

可視化 可視化の技術は、データセット（特に次元削減技術）（Van der Maaten and Hinton, 2008; Olah, 2014, 2015）、特徴量（Erhan et al., 2009; Olah et al., 2017）、重み、活性化（activations）（Carter et al., 2019）、構造（Reif et al., 2019）、神経回路網全体（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）を可視化する技術など、ニューラル構造の理解に役立つ。可視化の目的は、新たな詳細度（Olah et al., 2020）でニューラルネットワークを見ることである。

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).

摂動とアブレーション（Ablation） これらの手法は、相関関係ではなく、反事実（the counterfactual）を検証するように設計されている (Räuker et al., 2023)。摂動とは、モデルの入力を変更し、その出力の変化を観察する手法であり、アブレーション（切除）とは、ニューラルネットワークの一部をノックアウトする手法であり、ニューラルの活性化とネットワーク全体の行動との因果関係を立証するのに役立つ（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,

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

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).

パッチング パッチングとは、主要なコンポーネント（経路や活性化（paths and activations））を置き換えて、モデルの出力に対する反事実的な影響を解明する手法の総称である。その中でも活性化パッチングは、安全コミュニティでよく使われる手法である。活性化パッチングを適用し、同じニューラルネット上で正しい実行と破損した実行の両方を実施することで、研究者はモデル出力により重要なキーとなる活性化を特定することを目的としている（Nanda, 2023a）。実際には、パッチングは学習表現／概念のマッピングと編集に使用される。具体的なパッチング技術には、変換器におけるトークン表現の解釈（Li et al, 2021a; Bansal et al, 2021; Geva et al, 2021, 2022; Power et al, 2022; Olsson et al、 2022）、キー・クエリー積（key-query products）を研究して、トークンがどのように互いにアテンションするのかを理解する（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）、潜在空間における方向から意味のある学習された概念を識別する（概念から方向へ（Fong and Vedaldi, 2018; Kim et al., 2018）、方向から事後説明へ（Schneider and Vlachos, 2021））が含まれる。安全性とアラインメントの目的のために、これらの技法は特に欺瞞の検出に役立つ（Burns et al., 2022）

4.2.3 Outlook【展望】

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]

重ね合わせ（Superposition）はニューロンレベルでの分析を困難にする 重ね合わせとは、モデルが次元数よりも多くの特徴を表現する現象のことで、そのため特徴がニューロン（Arora et al., 2018; Olah et al., 2020; Elhage et al., 2022）に対応しなくなる。重ね合わせは、モデル内のすべての特徴を列挙することで AI の安全性を確保することを難しくする（Elhage et al, 2022; Nanda, 2023b）。Elhage et al. (2022) は、重ね合わせを解決する 3 つの方法を提案している：重ね合わせのないモデルを作成する（トレーニング時に対処する）、ニューラルネットに特徴がどのように格納されるかを記述する過剰な基礎（overcomplete basis）を見つける（事後的に対処する）、または両方のアプローチを混合する。注目すべきは、Bricken et al. (2023) が、グループ・ニューロンを解釈するためのスパース・オート・エンコーダ（sparse auto-encoder）を構築していることであり、重ね合わせを解決する有望な方向性を示している。

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)).

スケーラビリティ 前のセクションで述べたように、モデルの解釈可能性とその能力の間にはトレードオフが存在する（Alvarez Melis and Jaakkola, 2018）ため、性能を維持しながら実際のモデルを解釈することは、簡易的な試験モデル（toy models）にそれらの技術を適用するよりも難しくなる。したがって、スケーラビリティは、解釈可能性の研究者が解釈可能性に対するボトムアップアプローチ（機械論的解釈可能性）をとる場合に懸念となる。アテンション機構（Hudson and Manning, 2018）のようなトップダウン手法は、このようなボトルネックに直面することはないだろう。機械的解釈可能性研究では、技術のスケールアップ

[39] see Elhage et al. (2022) for details on conceptual and empirical research questions about superposition

（例えば、実モデルに回路解析を適用する（Wang et al., 2022））か、解析のスケールアップ（例えば、ニューラルネットワークでより大きな構造を見つける（Olah, 2023））のどちらかを望む。最終的には、ミクロな解析が、私たちが関心を寄せるマクロなモデルの行動に関する疑問（例えば、文脈内の学習能力（Olsson et al, 2022）や、計画のような高レベルの認知能力や欺瞞のような危険な能力（Anthropic, 2023b）に関するより多くの推測）に答えることを望んでいる。

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).

評価とベンチマーキング ベンチマーキングは、どのような方法が効果的かについて洞察し、その効率を定量化する、また、コミュニティの努力を明確で有意義な方向に導くことができる（Lipton, 2018; Casper, 2023; Krishnan, 2020; Mohseni et al., 2021; Madsen et al., 2022）。解釈可能性ベンチマークと基準（metrics）は、解釈可能性ツール（トロイの木馬の検出における有効性を評価する）（Casper et al., 2023a）、回路（特定の部分グラフが回路としてカウントされるかどうかをテストする）（Lawrence et al., 2023）、説明（説明の忠実性、包括性、十分性を検証する）（Lage et al., 2019; DeYoung et al., 2020; Krishna et al., 2022）を評価するために作られた。しかし、ある AI システムの内部論理（the inner logic）は、解釈可能性ツールが適用される前は未知であり（Samek et al., 2019）、異なる説明が互いに矛盾する可能性さえあるため（Neely et al., 2021; Krishna et al., 2022）、信頼できる評価ベンチマークや評価基準（metrics）を構築するのはかなり難しい（Krishna et al., 2022）。

4.3 Human Values Verification【人間的価値観の検証】

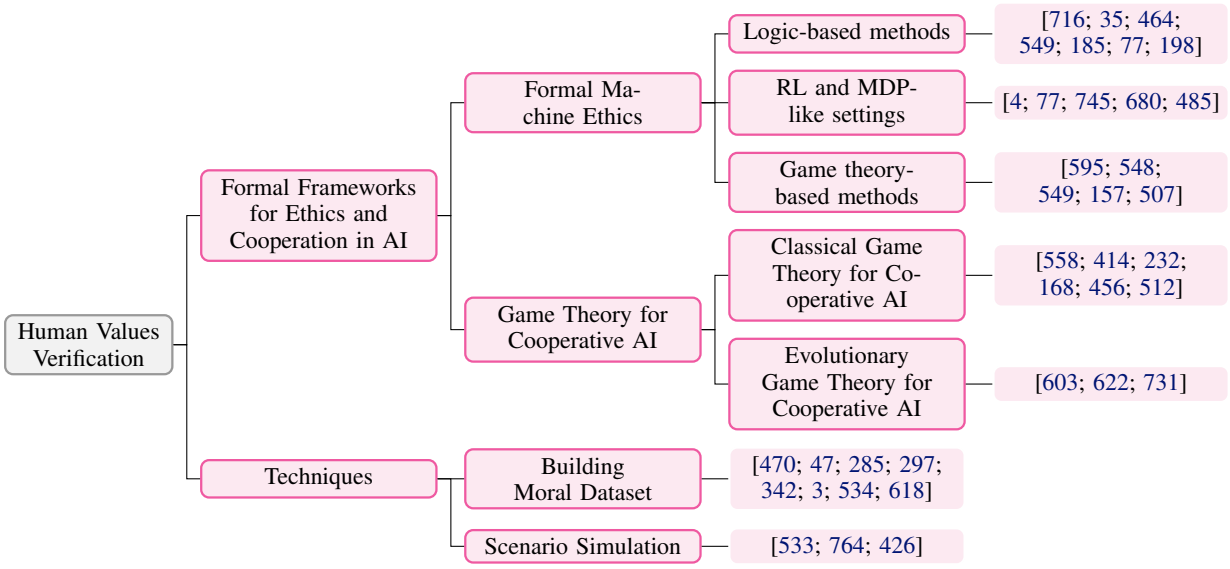

Figure 12: A Tree diagram summarizing the key concepts, logic, and literature related to Human Value Verification. The root of the tree represents Human Value Verification, which aims to *verify whether AI systems can adhere to the social norms and moral values*. The main branches represent the main 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.
図 12：人間的価値観の検証（Human Value Verification）に関する主要な概念、論理、文献をまとめたツリー図。ツリーのルートは人間的価値観の検証を表しており、AI システムが社会規範や道徳的価値観を遵守できるかどうかを検証することを目的としている。メインブランチは人間的価値観の検証の主な構造を表し、AI における倫理と協調のための形式的フレームワークや、価値観の検証の具体的な技法を含む。さらに枝分かれした部分には、それぞれの枝分かれを探求する主要な研究がリストアップされている。この図は、AI システムを人間的価値観や社会規範にアラインさせるための研究の方向性と具体的な技法の概要を示している。

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.

人間的価値観のアラインメントとは、AI システムがコミュニティの社会的・道徳的規範を遵守すべきであるという期待のことである（Chatila and Havens, 2019）。AI システムの能力が進歩するにつれて、AGI に近い能力を示すものも現れ始めている（OpenAI, 2023a）。将来的には、こうした AI システムに支配された自律エージェントが、私たちの日常生活に不可欠な存在になることが予想される（Lee et al., 2023b）。しかし、これらのシステムが人間的価値観に内在する複雑さと適応性を把握できなければ、その判断が社会的に否定的な結果をもたらす可能性がある。この文脈では、単に人間の意図にアラインするだけでは十分ではないかもしれない。したがって、AI システムと人間の間で、人間的道徳観や価値観のアラインメントを評価することが極めて重要になる（Weidinger et al., 2023）。このことは、より社会性を重視し、信頼性が高く、信用できる AI エンティティの設計の重要性を強調している。理論的研究と実践的技術の論理に従って、人間的価値観のアラインメントに関する議論をこの 2 つの側面に分けて行う： 人間 価値観のすり合わせの定式化 § 4.3.1 と評価方法 § 4.3.2 である。

4.3.1 Formulations【定式化】

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.

価値観の定式化は複雑であるため、アラインメントに関連する人間的価値観の側面をフォーマルに特徴付けるフレームワークを紹介する。具体的には、形式的な機械倫理と協調的 AI のためのゲーム理論という 2 つのトピックに焦点を当てる。前者では機械倫理の形式的フレームワークの構築に焦点を当て、後者ではマルチエージェントシステムの価値について議論する。

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:

形式的な機械倫理 1.2.3 節で初めて紹介した機械倫理（Yu et al., 2018; Winfield et al., 2019; Tolmeijer et al., 2020）は、倫理を順守する（ethically-compliant）AI システムを構築することを目的としている。ここでは、形式的フレームワークに焦点を当てた機械倫理の一分野を紹介する。形式的機械倫理のアプローチとして、論理ベース、RL/MDP ベース、ゲーム理論/計算社会的選択に基づく方法の 3 つを説明する：

- **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).

- **論理ベースの手法** 形式的機械倫理の主要な方向性の一つは、論理に焦点を当てている (Pereira et al., 2016b)。論理に基づく多くの研究は、機械倫理のために調整された特別な目的の論理システムを使用または提案している。例えば、ADC（Agent-Deed-Consequence）モデル（Dubljevic, 2020）、義務論理（deontic logic）（Von Wright, 1951; Arkoudas et al., 2005）、イベント計算（event calculus）とその変種（Berreby et al., 2017）などである。また、道徳的性質の形式的検証のための手法や、そのような形式的検証に対応する AI システムのフレームワークを開発する研究もある（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.

- **RL と MDP（Markov Decision Process：マルコフ決定過程）のような設定**。もう一つの研究は、MDP のような環境における計画のための統計的 RL や他の類似の手法に関するものである (Abel et al., 2016; Svegliato et al., 2021)。特に、いくつかの研究 (Wu and Lin, 2018; Svegliato et al., 2021) では、倫理指向の報酬関数を手動で設計することを利用している。逆に、他の研究 (Berreby et al., 2017; Murtarelli et al., 2021) では、倫理的意思決定と報酬機能の分離が追求されている。

- **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).

- **ゲーム理論に基づく手法** マルチエージェントの課題に対処するために、研究者はゲーム理論と計算的社会的選択に基づく機械倫理手法を開発してきた。Pereira et al. (2016a) によって提唱された、既存の研究の方法論は、Evolutionary Game Theory (EGT) (Pereira et al., 2016b)、古典的ゲーム理論 (Conitzer et al., 2017)、計算的社会的選択 (Rossi et al., 2011; Noothigattu et al., 2018) に大別される。

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).

協調的 AI のためのゲーム理論 協調的 AI（Dafoe et al., 2020, 2021）は、AI システムの非協力的で集団的に有害な行動（§1.1.2 参照）に対処することを目的としている。ここでは、§3.3.2 における MARL ベースの協調学習の紹介を補完するために、ゲーム理論に焦点を当てた協調的 AI の領域（the branch）を紹介する。この領域は、協調の能力を強調する MARL の傾向とは対照的に、協調のインセンティブを研究し、それを強化しようとする傾向がある。インセンティブの失敗の事例には、囚人のジレンマ（Phelps and Russell, 2023）やコモンズの悲劇（Perolat et al., 2017）のようなゲーム理論のジレンマがあり、協調能力の失敗の事例には、ロボットサッカーチームの悪い協調（Ma et al., 2022）がある。

- **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.

- **協調的 AI のための古典ゲーム理論** 協調的 AI の設定として、古典的ゲーム理論に焦点を当てた研究は数多くある。その中でも特に注目されているのが、シュタッケルベルグゲーム［先導者とされる寡占企業が価格決定した後に、追随者が価格決定を行う逐次手番ゲームの寡占モデル］、つまり、一人のプレーヤー（「リーダー」）が最初に動き、他のすべてのプレーヤー（「フォロワー」）がリーダーの動きに応じて動くゲームである。これは、ゲームにおけるコミットメント（すなわち、プレイヤーが優位に立つために特定の手や戦略にあらかじめコミットすること）をモデル化するのに適しており、Dafoe et al. (2020) によれば、コミットメントを理解することは、協力的 AI 研究の 4 つのピラーの 1 つである。シュタッケルベルグゲームに関する最近の研究としては、モデルへの境界合理性の導入（Pita et al., 2010）、動的モデル（Li and Sethi, 2017）、シュタッケルベルグ均衡の機械学習（Fiez et al., 2020）などがある。シュタッケルベルグゲーム以外にも、Dafoe et al. (2020) は、その現実性から、混合動機ゲーム（すなわち、純粋に協力的でもなく、純粋に競争的でもない一般的なゲーム）の研究の重要性を強調している。この点に関する最近の研究事例としては、合成集団における価値観の多様性と混合動機ゲームのパフォーマンスとの間に正の相関関係があることを発見した McKee et al. (2020) や、ゲーム結果のパレート的改善を誘導するために一般的なゲームのペイオフ行列（the payoff matrix）に介入を構成した Oesterheld and Conitzer（2022）などがある。

- **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.

- **協調的 AI のための進化ゲーム理論** Sachs et al. (2004) によって始められたもう一つの研究分野は、協調が進化からどのように生まれるかを理解することを目的としている。これには、ダーウィン的進化から生まれた人間の協調だけでなく、自己複製力学（the replicator dynamics）(Shususter and Sigmund, 1983) のような他の進化的設定の中で生まれる可能性のある AI システムにおける協調傾向も含まれる。これらの研究は、進化ゲーム理論（Weibull, 1997）と呼ばれる方法論を採用しており、多くの場合、動的シ

ステムのツールを用いて、他者とのゲームの結果によって繁殖の成功が決定されるエージェントの大規模な集団の長期的な進化的結果を研究している。この分野での最近の研究では、現実性を向上させるために、例えば、集団構造や戦略の複雑性コストなどの特徴をモデルに追加する傾向がある。

4.3.2 Evaluation Methods【評価手法】

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*.

本節では、アラインメントすべき適切な価値が既に得られていることを前提とする。しかし、それでもグッドハートの法則（Goodhart and Goodhart, 1984）のもと、複雑な人間的価値観を単純な報酬関数として定義することはできず、価値観アラインメントにも大きな課題をもたらす。具体的な人間的価値観アラインメントの手法を 3 つのパートに分けて紹介する： モラル・データセットの構築、シナリオ・シミュレーション。

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.

モラル・データセットの構築 モラル・アラインメントとは、AI システムがタスクを実行したり人間の意思決定を支援したりする際に、人間と互換性のある道徳的基準や倫理的ガイドラインを遵守することを指す（Min et al., 2023）。2018 年に開始された道徳的価値アラインメントの初期の試み（Awad et al., 2018）では、道徳的価値そのものの定義と評価が困難な問題であることが確認されている。このため、抽象的な道徳的基準（Hagendorff, 2022）や、多様なコミュニティ集団の平均的価値観によって駆動される様々な異なる基準（Awad et al., 2018）が登場し、道徳的価値のアシュアランスに関してさらに詳細な研究が進展している。

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.

道徳的価値のアシュアランスは通常、対応するデータセットを構築することで達成される。経験則（Rule-of-Thumb：RoT）は、人間社会でどのような行為が許容されるかを判断する尺度として機能する。このコンセプトに基づき、Emelin et al. (2021)、Forbes et al. (2020)、Ziems et al. (2022) は、それぞれ Moral Stories、SOCIAL-CHEM-101、Moral Integrity Corpus データセットを導入し、人間の社会的・道徳的規範の提供に焦点を当てている。Hendrycks et al. (2020) と Jin et al. (2022) は、それぞれ ETHICS と MoralExceptQA データセットを紹介し、現在のモデルが人間的価値観に倫理的に適合できないことを強調している。Abdulhai et al. (2022) は、モデルが特定のモラルや価値観を示す頻度が他のモデルよりも高いことを発見し、これらのモデルが示すモラルの基盤が人間のモラルの基盤とどのように関連しているかを明らかにした。Pan et al. (2023b) は、報酬と道徳的行動の間のトレードオフを探求し、両者の間にある種の緊張関係を発見した。

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.

シナリオ・シミュレーション シナリオ・シミュレーションは、データセットよりも複雑な形式であるため、実際の状況をより効果的に再現し、より良い結果を得ることができると考えられている。シナリオの形

式も様々である。Pan et al. (2023a) は、テキストアドベンチャーゲーム（text adventure games）を通じて、多様で道徳的に重要なシナリオ群を構築し、欺瞞、操作、裏切りなどの複雑な行動を評価した。一方、人間と機械の相互作用をシミュレートすることで、知的エージェントに人間的価値観を学ばせようとする研究もある。Yuan et al. (2022) は、人間と機械の間で双方向の価値観をアラインメントする方法を提案し、人間のフィードバックを通じて機械が人間の選好や暗黙の目的を学習することを可能にした。Liu et al. (2024a) は、AI をシミュレートされた人間社会のサンドボックス内に置き、AI が人間と社会の相互作用を模倣することによって、人間の社会的価値傾向を学習することを可能にした。

5 Governance【ガバナンス】

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.

AI システムの安全な開発とデプロイを保証するためには、技術的な解決策に加えて、ルールの作成と実施であるガバナンスが必要である。本セクションでは、AI ガバナンスの役割、AI をガバナンスする上での機能、ステークホルダー間の関係、効果的な AI ガバナンスに対するいくつかの未解決の課題を探ることで、AI ガバナンスに関する文献をサーベイする。

5.1 The Role of AI Governance【AI ガバナンスの役割】

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 ガバナンスの役割を探るには、ガバナンスを必要とする課題（the challenges）を特定しなければならない。社会の様々な分野への AI の導入と統合によって、様々な社会的・倫理的問題が生じうるし、既に生じている（AI Safety Summit, 2023）。例えば、AI アプリケーションは、不用意にも（inadvertently）社会の偏見を永続させ、人種差別やジェンダー差別をもたらす可能性がある（Caliskan et al., 2017; Perez et al., 2023）。さらに、こうしたシステムへの野放図な依存は、労働力の移動（Acemoglu and Restrepo, 2018）、社会経済的格差の拡大、独占的環境（monopolistic environments）の創出といった深刻な影響をもたらす可能性がある。

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.

AI システムは世界の安全保障を危険にさらす可能性を示している（Turchin and Denkenberger, 2020）。例えば、OpenAI の GPT-4 のシステムカード（system card）（OpenAI, 2023a）は、GPT-4 モデルの初期バージョンと、有用性と無害性を高めるためにファインチューニングされたバージョンが、他の危険な行動の中でも、偽情報、影響力のある作戦、新しい生化学物質の設計を可能にする能力を示すことを発見した。Urbina et al. (2022) はさらに、創薬モデルを反転させて 4 万個の有毒分子を生成することで、AI システムが合成生物学の悪用を可能にする可能性を示した。

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.

また、十分な安全策を講じなければ、人類に壊滅的あるいは実存的リスクさえもたらす可能性のある、さらにエージェント的で汎用的な AI システムが登場する見通しもある（McLean et al., 2023）。例えば、OpenAI の Weng (2023b) は、LLM のようなモデルが本質的に知的エージェントの脳として機能し、計画、内省、記憶、ツールの使用によって強化されると主張している。AutoGPT、GPT-Engineer、BabyAGI などのプロジェ

クトは、この進化を象徴している。これらのシステムは、複雑なタスクを自律的にサブタスクに分解し、人間の介入なしに決定を下すことができる。マイクロソフトの研究によれば、例えば GPT-4 は AGI の初期の兆候を示唆している（Bubeck et al., 2023）。このようなシステムが進化するにつれ、失業などの社会経済的影響が広範囲に及び、悪意ある行為者が有害な活動を行うためのツールを装備する可能性がある。

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.
AI ガバナンスの主な目的は、この多様なリスクを軽減することである。この目標を達成するために、関係者は各リスクのカテゴリーを十分に考慮し、バランスの取れた取り組みのポートフォリオを維持すべきである。

5.2 The Multi-Stakeholder Approach【マルチ・ステークホルダー・アプローチ】

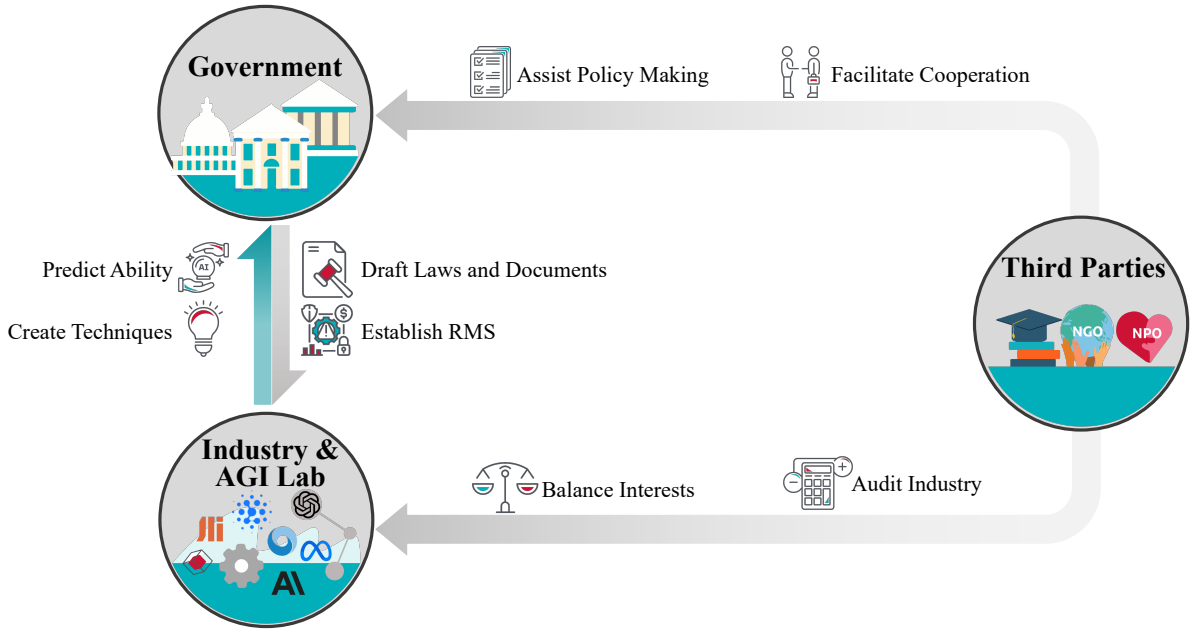

Figure 13: Our framework for analyzing AI governance at present. The proposed framework explains the nonexhaustive interrelationships and functions among three primary entities in AI governance: the government, industry and AGI labs, and third parties. The government's governance role encompasses regulating 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.
図 13：現在の AI ガバナンスを分析するための我々のフレームワーク。提案するフレームワークは、AI ガバナンスにおける 3 つの主要な主体（政府、産業界と AGI 研究所、第三者）の間の相互関係と機能を説明するものである。政府のガバナンスの役割には、産業界と AGI 研究所を規制し、政策文書を通じて将来の AI 開発の方向性（the trajectory）を指示することが含まれる。また、AI 関連の脅威を軽減するためのリスク管理システム（RMS）（Mannes, 2020）も考案する。産業界と AGI 研究所は、AI 開発に対する監視的予測を提供し、規制措置（モデル評価（Shevlane et al., 2023）など）を支援するための新たな技術的方法論を革新することで 応えている。サードパーティは、政府の政策立案のために専門家の助言を提供し、政府間の協力を促進するという二重の機能を果たす。産業界と AGI 研究所の関係では、これらの第三者は、情報の非対称性からくる競争の乱れを防ぐために、企業の利害を均衡させる手助けをする。また、独立した組織として、産業界や AGI 研究所に監査サービスを提供している。

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

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.
我々は、AI ガバナンスにおけるステークホルダー間の機能と関係を分析するためのフレームワークを提唱する（図 13 参照）。このフレームワークでは、3 つの主体について概説する。政府機関は、立法、司法、執行の権限を用いて AI 政策を監督し、国際協力にも関与する。産業界と AGI 研究所は、AI 技術を研究・開発し、ガバナンスのフレームワークの主体として、自らをガバナンスする技術を提案し、ガバナンス政策に影響を与える。アカデミア、NGO、NPO などの第三者は、コーポレート・ガバナンスや AI システム、その応用に関する監査だけでなく、政府の政策立案を支援する。

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.
マルチステークホルダーによる AI ガバナンスのあり方については、具体的な原則が提案されている。特に、Brundage et al. (2020) は、AI システムの安全性に関する主張をより検証可能なものにするために、制度、ソフトウェア、ハードウェアを導入することを主張している。

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.
政府 Anderljung et al.（2023）によれば、政府の規制には 3 つの構成要素が必要である：(1) 最先端 AI 開発者の適切な要件を決定するための標準的な開発プロセス、(2) 規制当局に最先端 AI 開発プロセスの進捗状況を把握させるための登録・報告要件、(3) 最先端 AI モデルの開発・展開における安全基準の遵守をアシュアランスするメカニズム。

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.
現在、欧州連合（EU）の AI 法（European Parliament, 2023）や米国の AI 法（Bipartisan Framework for U.S. AI Act, Blumenthal and Hawley, 2023）など、政府による規制や法律が世界規模で浮上している。このような規制は、AI システムの安全性とアラインメントのために不可欠である。

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)
産業界と AGI 研究所 産業界や AGI 研究所におけるガバナンスの取り組みは、AI システムのライフサイクル全体を通じて包括的な AI リスク評価を重視すべきである。Koessler and Schuett (2023); Schuett et al. (2023) の議論に基づくと、AI リスク評価の全サイクルは 5 つの段階から構成されると見ることができる。開発前リスクアセスメント、訓練前リスクアセスメント、およびデプロイ前リスクアセスメントはすべて、様々なツールを用いた影響とリスクの予測と分析を含むが、詳細さ、明確さ、および高度さが増していく（Koessler and Schuett, 2023）。デプロイ後のモニタリングは、モニタリングのメカニズムを確立し、デプロイ後にこれまでのすべての分析を継続的に更新する段階である（Koesler and Schuett, 2023）。外部からの監視には、バグ報奨金プログラム（Schuett et al., 2023）、外部からのレッドチーミング、第三者によるモデル監査（Schuett et al., 2023; Anderljung et al., 2023）などがある。

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 capa-

bilities 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).
AI システムに関連するリスクに対するセキュリティ対策を講じることは、AI 企業や関連実務者に広く受け入れられているようである。Schuett et al. (2023) は、AGI ラボが AI の安全性をアシュアランスするために、デプロイ前のリスク評価、危険な能力の評価、第三者によるモデル監査、モデル使用の安全制限、レッド・チーミングを実施することを、調査を受けた回答者の 98%がある程度または強く承認していることを示している。一方、Amazon、Anthropic、Google、Inflection、Meta、Microsoft、OpenAI などの大手 AI 企業は、セキュリティ対策を実施することを自主的に政府に約束している（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).
注目すべきは、多くの研究者が、安全性研究、リスク評価、規制準備のための時間を確保するために、高度な AI システムの開発の一時停止（pausing）を提案していることである（Bengio et al, 2023）。彼らの提案には、十分に高度なシステムすべてを包括的に一時停止すること（Bengio et al, 2023）や、特定の失敗モードに関する評価結果に応じて特定のクラスのモデルを条件付きで一時停止すること（Alaga and Schuett, 2023）、現在採用されている責任あるスケーリングポリシー（responsible scaling policy：RSP）の実践（Anthropic, 2023a）などが含まれる。

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).
第三者 Mökander et al. (2023) は、第三者監査の 3 つの重要な機能を提示している：(1) ガバナンス監査（LLM を設計し普及させる技術提供者の監査） (2) モデル監査（事前トレーニング後、リリース前の LLM の監査） (3) アプリケーション監査（LLM に基づくアプリケーションの監査）。

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).
METR は、当初アラインメント研究センター（ARC Evals, 2023; Kinniment et al., 2023）のプロジェクトであり、OpenAI と協力して GPT-4（OpenAI, 2023a）のレッド・チーミングを実施し、Anthropic と提携してクロード 2（Anthropic, 2023c）のレッド・チーミングを実施した。これらの取り組みには、毒性や偏見だけでなく、自律複製、操作、サイバーセキュリティ、生物兵器のリスクなどのフロンティア AI のリスクに関する評価も含まれている（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).
監査とは別に、第三者は政策立案を支援したり、国際的な協力を促進したりするなど、他の方法で AI ガバナンスをサポートすることができる（Ho et al, 2023）。例えば、Maas (2021) は、政府は技術に特化したルールよりも技術に中立的なルールを選ぶべきだと考えている。AI4People の「良い AI 社会のための倫理的フレームワーク：機械、リスク、原則、勧告（Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations）」 (Floridi et al., 2021) は、2019 年 4 月に発表された「信頼できる AI のための倫理ガイドライン（Ethics Guidelines for Trustworthy Artificial Intelligence）」 (Atomium-EISMD, 2023) に従っている。世界経済フォーラム（WEF）は、政府高官、協力団体、市民社会を集め、AI 分野における国際協力を促進することを目的に、パートナー団体と共同でグローバル AI アクションアライアンスを開始した（Kerry et al., 2021）

5.3 Open Problems【オープンという問題】

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.

既存の AI ガバナンスの分野には、数多くのオープンという問題が存在する。これらの問題には明確な答えがないことが多く、これらの問題を議論することで、より良いガバナンスを促進できることが多い。効果的な AI ガバナンスのために、我々は主に国際ガバナンスとオープンソース・ガバナンスについて議論し、議論を通じて AI の安全な発展を促進することを望んでいる。

5.3.1 International Governance【国際的ガバナンス】

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.

AI 技術の急速な進展と世界的な普及の中で、AI の国際的ガバナンスの必要性が高まっている（Summit, 2023）。AI ガバナンスのグローバルなフレームワークを確立する必要性、その規範性と正当性を確保する手段（Erman and Furendal, 2022）など、重要な議論が展開されている。これらのテーマは、ますます詳細で複雑なレベルへと変化していく。また、［2023 年］7 月の安全保障理事会でアントニオ・グテーレス国連事務総長が述べたように、生成 AI は、規模の大小にかかわらず、プラスとマイナスの両方の影響をもたらす巨大な可能性を秘めており、AI のリスクを軽減するための行動を取らないことは、現在と将来の世代の福祉を守る義務を著しく怠ることになり（Guterres, 2023）、 したがって国際ガバナンスは世代間にも影響を及ぼす。そこで、本セクションでは、国際的な AI ガバナンスの重要性と実現可能性を、グローバルな破局的 AI リスクの管理、AI の機会の管理、歴史的な取り組みと現在の取り組みという 3 つの側面から、世代間と世代間の両方の視点を交えて検討する。国際的な AI ガバナンスの構造を展望する上で、革新的な考え方に貢献することを目指す。

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).

グローバルな破局的 AI リスクの管理 AI 技術の絶え間ない進歩は、世界の発展と繁栄に計り知れない可能性を約束している（Vinuesa et al., 2020）。しかし、その裏には必然的にリスクが潜んでいる。市場における野放図な競争や地政学的な要因によって、高度な AI システムの開発やデプロイが早まる可能性があり、その結果、負のグローバル外部性が生じる（Tallberg et al., 2023）。AI システムに組み込まれた人種や性別の偏見（Swaugerarchive, 2020）といった既存の不平等が増幅されることで、世代間の倫理的差別（intergenerational ethical discrimination）が生じる可能性もある。こうしたリスクは国際的かつ世代間的なものであるため、国際的なガバナンスの介入によって、こうした破局的なグローバルな AI の課題を軽減することができると思われる。例えば、国家間のコンセンサスは、潜在的な AI の軍拡競争を緩和するのに役立ち、一方、業界全体の合意は、高度な AI システムの性急で無責任な開発を回避し、AI の長期的かつ持続可能な発展を確保することができる（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.

AI の機会を管理する　AI の発展によってもたらされる機会は均等に分配されないため、地域間のデジタル不平等が永続化し、AI 開発の持続可能性が損なわれる可能性がある。AI の発展における地理的なばらつきは、その経済的・社会的便益の不公平な分配を示唆しており、発展途上国や特定のグループがこうした便益から排除される可能性がある (Ho et al., 2023; Tallberg et al., 2023)。さらに、テクノロジー分野での意思決定権が限られた個人に集約される（Sara Stratton, 2021; Noble et al., 2021）ことで、世代間の影響が生じる可能性もある。このような利益配分の不平等は、国際的なガバナンスを通じて緩和することができる。AI の普及、教育、インフラ整備によって促進される AI の機会配分に関する効果的な国際的コンセンサスと協調（Opp, 2023）は、AI から得られる利益のバランスの取れた配分を保証し、その継続的発展における持続可能性を促進することができる。

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.

歴史と現在の取り組み　AI 技術が急増する以前から、国際社会は影響力のある技術や重要事項を協調的に規制する枠組みを定めていた。例えば、気候変動に関する政府間パネル（IPCC）は、気候環境問題を評価するために専門家を招集し、科学的コンセンサスを醸成した（Ho et al., 2023）。国際民間航空機関（ICAO）は、国際的な規制を標準化し監督すると同時に、加盟国がこれらの法律を遵守しているかを評価した（Ho et al., 2023）。国際原子力機関（IAEA）は、その世界的な広がりと洗練された監視・評価メカニズムによって、原子力の調和のとれた利用を推進した。現在では、複数の国際機関が AI ガバナンスに関するコンセンサスに達している。2019 年、G20 加盟国は人間中心の人工知能原則に焦点を当てた閣僚宣言をまとめた（G20, 2019）。同時に、経済協力開発機構（OECD）は人工知能に関する OECD 原則を定めた（OECD, 2019）。IEEE スタンダード・アソシエーションは、自律的でインテリジェントなシステムの設計と実装に関与するすべてのステークホルダーが、倫理的な懸念を強調する適切な教育、訓練、動機付けを受け、それによって人類の向上のためにこれらの技術を進歩させることを確保することを目的とした世界規模のイニシアチブを開始した。(Chatila and Havens, 2019)。2021 年、国連教育科学文化機関 (UNESCO) は、AI 倫理に関する史上初の世界標準（global standard）を作成した (UNESCO, 2021)。この標準は、AI システムを人類と社会のために機能させるための基礎を築き、AI システムのコントロールを失うことによって引き起こされる潜在的な危害を防止することを目的としている。2023 年には、英国ロンドンで AI 安全サミットが開催された。各国は AI のリスクと機会について円卓会議を行い、共同でブレッチリー宣言を発表した（Summit, 2023）。また、国際 AI 機構（IAIO）のような、AI に関する国際的なガバナンスの枠組みも提案されている

（Trager et al., 2023）。これらの先例や研究成果が、将来、AI ガバナンスのための弾力的で長期的な国際的フレームワークを開発するための刺激となり土台となることを願っている。

5.3.2 Open-Source Governance【オープン・ソース・ガバナンス】

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.

現代の AI モデルのオープンソース化をめぐる議論は、AI ガバナンスにおいて、特にこれらのモデルがより強力になるにつれて、論争の的となっている（Seger et al.、2023）。これらのモデルをオープンソース化することに関連する潜在的なセキュリティ上の危険性は、AI 研究者や政策立案者の間で引き続き議論の核心となっている。オープンソースの AI ガバナンスにおける攻撃と防御のバランスについても、依然として議論が続いている。オープンソースモデルがモデルの安全性を高めるのか、それとも悪用のリスクを高めるのかについては、いまだに議論が続いている。Shapiro and Siegel (2010) で言及されているように、情報公開の有効性は、潜在的な攻撃者がすでに知識を保有している可能性と、透明性を新たな脆弱性の特定と解決に転換する政府の能力に依存する。Weng（2023a）が敵対的攻撃について論じているように、AI 分野における攻撃と防御のバランスについて、すでに予備的な議論を行っている学者もいる。AI システムにおいて攻撃と防御の適切な均衡が築けなければ、オープンソース化は AI システムの悪用という重大なリスクを生じさせる可能性がある。

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.

Seger et al. (2023) が述べたオープンソースモデルの定義、すなわち、モデルのアーキテクチャーとウェイトをオープンに公開し、誰でも修正、研究、さらなる開発、利用ができるようにすることである。現在、最も認知されているオープンソースモデルには、Llama2、Falcon、Vicuna などがある。このセクションでは、オープンソースモデルがもたらすセキュリティ上の利点と潜在的な脅威を評価し、これらのモデルのオープンソース化の実現可能性についての議論を促進する。最終的に、我々の目的は、既存の研究からの洞察を統合し、これらのモデルの安全な実装を確認するための将来のオープンソースの方法論に対する提案を提示することである。

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:

オープンソース化の議論 既存のモデルのオープンソース化を支持する見解は、この方法が、いくつかの点で、これらのモデルに内在するセキュリティリスクを軽減できることを示唆している：

- **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).

- **モデルのセキュリティを強化する** Meta 社の Llama2 のリリースブログ（Meta, 2023）での主張は、これにより開発者や技術コミュニティがモデルのテストを実施できるようになるという信念を表している。その結果、問題の迅速な特定と解決によって、モデルの安全性を大幅に強化することができる。これとは対照的に、別の観点からは、既存のモデルをオープンソース化することで、関連するリスクの

認識が深まり、潜在的な危険に対するより大きな関心、調査、緩和が促進されることが示唆されている（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.
- **権力と統制の分権化（Decentralization）を促進する**　オープンソース化は、経済的、社会的、政治的な領域を含む様々な分野において、主要な AI 研究所の支配力を低下させる効果的な戦略として広く認識されている（Seger et al., 2023）。その事例は、Stability 社が Stable Diffusion をオープンソース化した中核的な理由に明示されている：彼らは、中央集権的で選挙で選ばれたわけでもない組織が AI 技術をコントロールするのとは対照的に、個人やコミュニティに信頼を置いているのだ（Mostaque, 2022）。さらに、ある論者はオープンソースと啓蒙時代のアナロジーを提唱し、分散型管理は人類と社会の力と善に対する信頼を強化すると主張する（Howard, 2023）。

Arguments against Open-sourcing　Critics of open-source models assess the potential for misuse from the following viewpoints:
オープンソースに対する反論　オープンソースモデルを批判する人々は、以下の観点から悪用の可能性を評価している：

- **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).
- **有害な事例（Detrimental Instances）へとファインチューニングされる可能性**　現在の研究では、化学や生物学における毒性を緩和するという当初の設計意図とは相反する AI システムが、新たな化学毒素（Urbina et al, 2022）や生物兵器（Sandbrink, 2023）を製造する可能性を秘めていることが厳然と確認されている。このようなモデルの悪意あるファインチューニングは、重大なセキュリティ・リスクの顕在化につながる可能性がある。その上、言語モデルは、いったんファインチューニングされれば、熟練した執筆者を模倣し、説得力のある偽情報を作り出すことができる（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.
- **システムの脱獄（Jailbreaks）を不用意に助長する**　研究によれば、オープンソース化されたモデルウェイトに自由にアクセスすることで、システムのセキュリティー対策を回避しやすくなる可能性が指摘されている（Seger et al., 2023）。Zouet al. (2023b) は、Vicuna- 7B と 13B を使って攻撃用接尾辞（suffixes）を開発することで、この可能性を示した。ChatGPT、Bard、Claude のような、容易にアクセス可能なインターフェースに実装されると、これらは望ましくない生成（unwanted generations）を引き起こした。したがって、モデルをオープンソース化することは、オープンソース化されていないモデルの保護プロトコルを意図せずして弱体化させ、結果としてモデルの悪用の可能性を増大させるかもしれない。

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.

オープンソース・ガバナンスに関する暫定的な結論 AI モデルのオープンソース化に関する議論は、現在のところ、依然として未確定だが、AI モデルの開示（disclosure）が大きなリスクをもたらすことはないという見解が優勢である。我々の論考は、このトピックに関する既存の視点を統合するだけでなく、より高度な AI システムをオープンソース化することの慎重さを考慮した将来の議論の基盤を準備するものである。

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).

高度な AI システムをオープンソース化するための既存のガイドラインには、ファインチューニングや段階的なモデル公開によって悪用される可能性を定量化し、リスクを評価するなどの対策が含まれている（Solaiman et al., 2019; Seger et al., 2023）。一方、政策立案者は、これらのオープンソースモデルに対する厳格なコンプライアンスプロトコルを確立しつつある。例えば、欧州の政策立案者は、モデルが「性能、予測可能性、解釈可能性、相関性、安全性、サイバーセキュリティをライフサイクル全体にわたって」持つべきであると主張している（Chavez, 2023）。

5.4 Rethinking AI Alignment from Socio-technical Perspective【社会技術的観点からの AI アラインメント再考】

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].

これまでの議論では、AI アラインメントの中核となる AI システムに主眼を置いてきた。私たちは、システムのライフサイクル全体を通して、人間の意図や価値観とシステムをアラインさせるための戦略について、フォワード・アラインメントとバックワード・アラインメントの両方を考慮しながら検討する。将来、AI はより困難でリスクの高い意思決定、例えば「公平性を保つために資源をどのように配分するか」「どの医薬品を承認すれば安全か」などを扱うようになるだろう。このような意思決定には、十分な情報に基づいた回答を得るための高度な専門知識が必要とされるだけでなく、価値判断も含まれるため、情報に精通した個人の間で、異なる価値観に基づく強い意見の相違が生じることになる。さらに、AI システムは誤った価値観を伝達し、世論を揺さぶり、文化的侵略を促進し、社会分裂を悪化させるかもしれない（Goldstein et al., 2023）。かつてシンガポール AI 会議（SCAI）は、人類が繁栄するために世界の AI コミュニティが取り組むべき課題を総合的に定式化した 12 の質問を紹介したことがある。

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.

アラインメントの分野において、私たちは次のような問いに関心を持っている。: AI システムが社会技術的な存在へと進化していく中で、アラインメント技術はどのように人間社会にもたらす課題を軽減することができるのか？具体的には、アラインメント技術を通じて AI システムへの価値観の組み込みを探求し、セキュリティ手法に関する洞察を提供する。また、将来の AI システムがもたらす社会技術的課題に対処するために必要なアラインメント技術を明らかにすることを目指す。

5.4.1 How to incorporate value into AI systems?【どのように AI システムに価値を組み込むか？】

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

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

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.

AI システムを人間のモラルや社会的価値とアラインさせることは、アラインメント技術の重要な目的である。しかし、現在の技術（例えば、RLHF）は、特定の価値を区別することなく、主に選好をブレンドし、人間の選好のみに焦点を当てている。人間の選好は基本的なアラインメントの問題に効果的に対処する。つまり、モデルが人間の意図や安全性にアラインすることをアシュアランスするが、モラルや社会的価値にはアラインしない。しかし、将来の AI システムの重大な問題における些細なミスは、見解の異なる人々の間で意見の相違を引き起こす可能性がある。人間的価値観を真に理解することは、AI システムが様々なシナリオやイデオロギーを超えて一汎化し適応するために極めて重要である。AI システムに価値観を組み込むには、一般的に、個人の価値観にアラインさせること（4.3 節）と、集団の価値観にアラインさせることの 2 つの側面がある。

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.

このパートでは、主に第二のトピックについて論じる。集団的価値観のアラインメントにおける主な課題は、どの集団を含めるかを決定することにある。一般的なアプローチは、無知のベールに代表されるように、公平、正義、利他主義といった普遍的な価値を定義することである。しかし、この作業は理論的なものにとどまっている。別のアプローチでは、普遍的な価値の定義を避け、その代わりに文化間の価値の最も広い重複を求める。Bakker et al. (2022) は、様々な属性から選好を集め、言語モデルを訓練し、多様な社会厚生関数を用いて結果を集計することで、このアプローチを開始した。同様に、意思決定を強化するために、シミュレートされた熟議民主主義が提案されている（Leike, 2022）。具体的には、多様な集団（demographics）からなる個人が、AI の支援を受けながら、価値観の異なるトピックについてコンセンサスを得る。このデータが新たなモデルのトレーニングに反映され、新たな価値観の問題により適切に対応するための熟議民主主義のシミュレーションが可能になる。

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.

さらに、すべてのユーザーに対してコンセンサスとなる答えを提供するのではなく、集団的な価値観のアラインメントを行うことで、AI システムが特定の人口統計グループに合わせた回答をするように促すべきである。言い換えれば、特定の質問や特定の対話において、どのような価値観がモデルの応答を導くべきなのだろうか？民主的ファインチューニング（MAI, 2023）は、価値カードとモラルグラフを使用して様々な価値観をリンクさせ、ファインチューニングされた LLM が応答する前にモラルの文脈を振り返ることを可能にする。

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.

しかし、多くの価値観の議論は静的な価値観を前提としているが、社会的価値観は実際には動的で進化している。価値観をアラインした AI システムが、環境の価値観の変化にどのように動的に適応できるかを探ることは極めて重要である。フューチャー・テクノロジーは、まず静的な価値観のアラインメントに取り組む必要があり、これにはアラインメントのために人間集団をサンプリングする戦略も含まれる。Bakker et

al. (2022) は、あるサブグループで黙々と合意形成された声明が、除外されたメンバーの反対を招くことを発見し、コンセンサスが個人の意見に敏感であることを強調した。国際協力のためには、共有データセンターの設立が必要であるが、どの文明を含めるか、その価値観がアラインできるかどうかをまず決定する必要がある。

5.4.2 How to use Alignment techniques to support AI Governance?【AI ガバナンスをサポートするため、どのようにアラインメント技術を使用するか？】

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.

AI システムが実世界の様々な意思決定シナリオに採用されるにあたり、その信頼性と信用性を確保することは極めて重要である。一方では、言語モデルは依然として使用中に錯覚を示し、他方では、システムの信頼性は、個々のテスト環境下におけるシステムの信頼性と、人間との相互作用における信頼性の 2 つの部分から構成される。もう一つの課題は、ユーザーに観察可能で説明可能な意思決定プロセスを持つシステムを構築することである。社会的な観点からは、分野横断的な AI システムの普及も潜在的なリスクとなる。このリスクは、下流工程での応用（downstream applications：現場での活用の意）を考慮することなく技術の進歩に注力しがちな AI 開発者と、安全性への十分な配慮や再現可能な成功例の検証なしに AI システムを各分野に導入しかねない AI 導入者との間のギャップから生じる。したがって、AI 採用者がモデルの有用性と適切性を正確に評価し、AI 規制当局が AI システムのリスクを迅速に特定して安全性アラート（safety alerts）を発することができるような枠組みを構築することが極めて重要である。

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.

アラインメント技術は、AI システムの同期化、独立化、厳格な評価を促進することができる。AI 開発者は、社会経済的、文化的、その他の違いの重要性を認識し、学習プロセスにおける適切なバイアス処理を優先すべきである。さらに、AI システムを監査するための強固で公正な評価方法とデータセットの開発を目指すべきである。Zhu et al. (2023) は、大規模言語モデルのための初の動的テストプロトコルを提案し、Directed Acyclic Graphs (DAG) を利用してテストデータを動的に生成することで、テストデータの記憶や汚染のリスクを低減している。さらに、新しい堅牢なセキュリティプロトコル評価手法も導入されている：Shlegeris と Greenblatt(2023) は、危険なほど強力で欺瞞的なモデルを管理するための敵対的ポリシーの構築を提案し、Greenblatt et al.(2023) は、モデルの有害性と欺瞞性のレベルに基づいてモデルを監督するための（非）信頼された編集（(un)trusted editing）を提案している。また、今後の取り組みとして、AI システムが評価システムを悪用した報酬ハッキングを防止し、AI 規制当局に説明可能で独立した一元的な評価システムを提供することを目指すべきである。

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.

AI を導入する企業や業界は、ユースケースを徹底的に評価し、成功例と失敗例の両方を紹介するケーススタディを共有するために、資金と計算資源を割り当てるべきである。同様に重要なのは、導入する企業に対する下流工程での応用に関するトレーニングである。

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

6 Conclusion【結論】

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*.

この調査では、人間の意図や価値観に沿った振る舞いをする AI システムを構築することを目的とする、AI アラインメントについて広範に紹介した。アラインメントの目的は、堅牢性（Robustness）、解釈可能性（Interpretability）、制御可能性（Controllability）、倫理性（Ethicality）（RICE）であり、アラインメント手法の範囲は、フォワード・アラインメント（アラインメント訓練によって AI システムをアラインメントさせる）とバックワード・アラインメント（システムのアラインメントの証拠を得て、ミスアラインメントのリスクを悪化させないように適切に制御する）から構成される。現在のところ、フォワード・アラインメントにおける 2 つの注目すべき研究分野は、フィードバックからの学習と分布シフト下での学習であり、バックワード・アラインメントはアシュアランスとガバナンスで構成されている。

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.

アラインメントが他の多くの分野と一線を画しているのは、その多様性である（Hendrycks, 2022）。アラインメントは、複数の研究の方向性と方法が緊密に組み合わさっており、方法論の共有とは対照的に、共有された目標によって結びつけられている。この多様性はメリットをもたらす。異なる方向性が互いに競い合い、ぶつかり合うことで、アイデアの交配が生まれ、イノベーションが促進される。これはアラインメントサイクル（図 2 参照）に反映されており、4 つのピラーは AI システムのアラインメントを継続的に改善する自己改善ループに統合されている。一方、このような研究の方向性の多様性は、この分野への参入障壁を高めており、新規参入者にも経験者にも役立つ、よく整理された調査資料の編纂を義務付けている。このサーベイでは、アラインメントに関する包括的かつ最新の概要を提供することで、このニーズに応えようと試みている。

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.

我々は、アラインメントを広範かつ包括的な特徴として導入することで、この分野における多様性を完全に説明しようと試みている。アラインメントに関する我々の調査は、この分野における主要な研究課題のほとんど全てにスポットライトを当て、またアシュアランスやガバナンスの面における現実の実践にもスポットライトを当てている。我々は、アラインメントの境界はしばしば曖昧であり、議論の対象となることを認識している。したがって、RICE の原則を提案する際には、アラインメントの広範な特徴を明確な選択肢として提示する。

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.

一方、このような調査は、継続的に見直され、更新される長期的な取り組みである必要があると我々は認識している。アラインメントの問題も方法も、機械学習の発展に密接に付随している。この速いペースでの発展は、新しい資料やフレームワークがわずか数年で時代遅れになることを意味する。この事実は、私たちが最新の開発を反映させるためにサーベイを作成し、継続的なメンテナンスと更新を義務付けている理由の 1 つである。

We conclude the survey by looking ahead and presenting the key traits in this field that we believe ought to be preserved or fostered.
最後に、私たちがこの分野で守るべき、あるいは育むべきと考える重要な特性を提示し、このサーベイを締めくくる。

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).
新しい課題とアプローチのオープンエンドな探求 アラインメントに関する言説の多くは、LLM や大規模ディープラーニングにおけるその他のブレークスルーが開発される以前の古典的な研究に基づいている。したがって、機械学習分野でこのようなパラダイムシフトが起こると、アラインメントにおけるいくつかの課題が目立たなくなる一方で、他の課題がより注目されるようになるのは当然である。より重要なことは、機械学習の方法論におけるこのシフトと、AI システムの社会への統合がますます厳しくなる広範な傾向（Abbass, 2019）は、以前には想定できなかった新たな課題をもたらすということである。このため、これまで軽視されていた新たな課題を積極的に探し出し、オープンエンドな探求に取り組む必要がある。さらに、このような探求は課題に限定される必要はなく、アプローチやソリューションに関しても同様の考え方を導入し、質問と答えの両方についてより多様なポートフォリオを構築する必要がある（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).
未来志向と現在志向の視点の組み合わせ アラインメントは、現在のシステムよりも強力な能力を持つ潜在的な高度 AI システムによる危害を強調してきた（Ngo, 2020a）。このようなシステムは、はるか未来に誕生するかもしれないし、あるいは数年先のことかもしれない（Stein-Perlman et al, 2022）。前者の可能性は、外挿トレンドや仮説的シナリオを検討する必要がある（Carlsmith, 2022）。対照的に、後者の可能性は、現在のガバナンス機構と連携し、現在のシステムをより高度なもののプロトタイプとして使用する、現場での取り組みの必要性を強調している（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).
政策適合性の重視 アラインメント研究は、研究者、産業界関係者、政府、非政府組織などの参加によるエコシステムの中で行われている。したがって、AI のアラインメントと安全性のエコシステムのニーズに応える研究は有用である。そのようなニーズには、様々なガバナンス・スキームに対する主要な障壁の解決、例えば、極限的リスク評価（Shevlane et al., 2023）、コンピューティング・ガバナンスのためのインフラストラクチャー、AI システムについて検証可能な主張を行うためのメカニズム（Brundage et al., 2020）などが含まれる。

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.
社会的複雑性と道徳的価値の重視 AI システムの社会への統合が進むにつれ（Abbass, 2019）、アラインメントは単一エージェントの問題ではなく、社会的な問題となる。ここで、社会的という意味は 3 つある。

[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).

1. 複数の AI システムと多人数の人間との相互作用を特徴とするマルチエージェント設定におけるアラインメント研究（Critch and Krueger, 2020）。

2. 人間の道徳的・社会的価値をアラインメントに組み込むこと（1.2.3 節と 4.3 節参照）。これは機械倫理と価値アラインメントの分野に密接に関連している（Gabriel, 2020; Gabriel and Ghazavi, 2021）。

3. AI システムが社会に与える影響のモデリングと予測。これには、社会科学からのものを含め、社会システムの複雑性にアプローチする手法が必要である。潜在的に有用な方法論の事例としては、社会シミュレーション（Bonabeau, 2002; De Marchi and Page, 2014; Park et al., 2023a）やゲーム理論（Critch and Krueger, 2020）などがある。

Acknowledgments

We thank David Krueger, Anca Dragan, Alan Chan, Stephen Casper, Haoxing Du, and Lawrence Chan for their helpful and constructive feedback on the manuscript. We thank Yi Qu for the graphical design and refinement of the figures in our survey.

David Krueger 氏、Anca Dragan 氏、Alan Chan 氏、Stephen Casper 氏、Haoxing Du 氏、Lawrence Chan 氏の原稿に対する有益かつ建設的なフィードバックに感謝する。また、本調査の図表をデザインし、洗練させてくれた Yi Qu 氏に感謝する。

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.

[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.

[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.

[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-Raphael 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/article/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.

[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/HBxe6wdjxK239zajf/what-failure-looks-like.

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

[147] Paul Christiano. 2023. Thoughts on the impact of rlhf research. https://www.lesswrong.com/posts/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-technical-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/posts/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.org/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.alignmentforum.org/posts/PZtsoaoSLpKjjbMqM/the-case-for-aligning-narrowly-superhuman-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/pRkFkzwKZ2zfa3R6H/without-specific-countermeasures-the-easiest-path-to.

[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.google/technology/safety-security/googles-ai-red-team-the-ethical-hackers-making-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.

[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.

[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*.

[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.alignmentforum.org/posts/rxoBY9CMkqDsHt25t/eliciting-latent-knowledge-elk-distillation-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/2023-11-07-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.

[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*.

[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.

[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.

[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 Ordónez, 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.

[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.

[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/EeAgytDZbDjRznPMA/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. https://cset.georgetown.edu/publication/key-concepts-in-ai-safety-interpretability-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-robustness-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] 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.

[604] Dorsa Sadigh, Anca D Dragan, Shankar Sastry, and Sanjit A Seshia. 2017. *Active preference-based learning of reward functions*. Escholarship.

[605] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. 2020. Distributionally robust neural networks. In *International Conference on Learning Representations*.

[606] 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.

[607] 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*.

[608] 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.

[609] Jonas B Sandbrink. 2023. Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools. *arXiv preprint arXiv:2306.13952*.

[610] 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.

[611] 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.

[612] 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.

[613] Fumihiro Sasaki and Ryota Yamashina. 2020. Behavioral cloning from noisy demonstrations. In *International Conference on Learning Representations*.

[614] 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*.

[615] 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.

[616] Stefan Schaal. 1996. Learning from demonstration. *Advances in Neural Information Processing Systems*, 9.

[617] Stefan Schaal. 1999. Is imitation learning the route to humanoid robots? *Trends in cognitive sciences*, 3(6):233–242.

[618] Nino Scherrer, Claudia Shi, Amir Feder, and David Blei. 2024. Evaluating the moral beliefs encoded in llms. *Advances in Neural Information Processing Systems*, 36.

[619] 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.

[620] 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*.

[621] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

[622] Peter Schuster and Karl Sigmund. 1983. Replicator dynamics. *Journal of theoretical biology*, 100(3):533–538.

[623] 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.

[624] Shalom H Schwartz. 1994. Are there universal aspects in the structure and contents of human values? *Journal of social issues*, 50(4):19–45.

[625] 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*.

[626] 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.

[627] Amartya Sen. 1986. Social choice theory. *Handbook of mathematical economics*, 3:1073–1181.

[628] Egemen Sert, Yaneer Bar-Yam, and Alfredo J Morales. 2020. Segregation dynamics with reinforcement learning and agent based modeling. *Scientific Reports*, 10(1):11771.

[629] 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*.

[630] Rohin Shah and Vikrant Varma. 2022. More examples of gmg. https://www.alignmentforum.org/posts/Cfe2LMmQC4hHTDZ8r/more-examples-of-goal-misgeneralization.

[631] 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*.

[632] 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.

[633] Jacob N Shapiro and David A Siegel. 2010. Is this paper dangerous? balancing secrecy and openness in counterterrorism. *Security Studies*, 19(1):66–98.

[634] 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*.

[635] 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*.

[636] Kenneth Shaw, Shikhar Bahl, and Deepak Pathak. 2023. Videodex: Learning dexterity from internet videos. In *Conference on Robot Learning*, pages 654–665. PMLR.

[637] 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*.

[638] 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.

[639] 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*.

[640] 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.

[641] Daniel Shin, Anca D. Dragan, and Daniel S. Brown. 2023. Benchmarks and algorithms for offline preference-based reward learning. *Trans. Mach. Learn. Res.*, 2023.

[642] 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.

[643] Buck Shlegeris and Ryan Greenblatt. 2023. Meta-level adversarial evaluation of oversight techniques might allow robust measurement of their adequacy. In *Alignment Forum*.

[644] 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.

[645] 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*.

[646] 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.

[647] 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.

[648] David Silver, Satinder Singh, Doina Precup, and Richard S Sutton. 2021. Reward is enough. *Artificial Intelligence*, 299:103535.

[649] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

[650] 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*.

[651] Munindar P Singh. 2014. Norms as a basis for governing sociotechnical systems. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(1):1–23.

[652] Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. 2022. Defining and characterizing reward gaming. *Advances in Neural Information Processing Systems*, 35:9460–9471.

[653] 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.

[654] Nate Soares. 2018. The value learning problem. In *Artificial intelligence safety and security*, pages 89–97. Chapman and Hall/CRC.

[655] Nate Soares and Benja Fallenstein. 2014. Aligning superintelligence with human interests: A technical research agenda. *Machine Intelligence Research Institute (MIRI) technical report*, 8.

[656] Nate Soares, Benja Fallenstein, Stuart Armstrong, and Eliezer Yudkowsky. 2015. Corrigibility. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*.

[657] 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.

[658] Emily H. Soice, Rafael Rocha, Kimberlee Cordova, Michael Specter, and Kevin M. Esvelt. 2023. Can large language models democratize access to dual-use biotechnology?

[659] 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*.

[660] Jiaming Song, Hongyu Ren, Dorsa Sadigh, and Stefano Ermon. 2018a. Multi-agent generative adversarial imitation learning. *Advances in Neural Information Processing Systems*, 31.

[661] Yang Song, Rui Shu, Nate Kushman, and Stefano Ermon. 2018b. Constructing unrestricted adversarial examples with generative models. *Advances in Neural Information Processing Systems*, 31.

[662] Giovanni Spitale, Nikola Biller-Andorno, and Federico Germani. 2023. Ai model gpt-3 (dis)informs us better than humans. *Science Advances*, 9(26):eadh1850.

[663] 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*.

[664] 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.

[665] 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)*.

[666] 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.

[667] 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.

[668] 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.

[669] 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.

[670] 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.

[671] 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.

[672] 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.

[673] 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.

[674] 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.

[675] 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.

[676] 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.

[677] 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.

[678] 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.

[679] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.

[680] 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.

[681] 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.

[682] 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.

[683] 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*.

[684] 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*.

[685] 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.

[686] Ming Tan. 1993. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the tenth international conference on machine learning*, pages 330–337.

[687] 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.

[688] Annalisa T Taylor, Thomas A Berrueta, and Todd D Murphey. 2021. Active learning in robotics: A review of control principles. *Mechatronics*, 77:102576.

[689] 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/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai.

[690] 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.

[691] 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.

[692] 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*.

[693] 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.

[694] 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.

[695] 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.

[696] 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*.

[697] 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*.

[698] 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.

[699] Grigorios Tsoumakas and Ioannis Katakis. 2007. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13.

[700] Alexey Turchin and David Denkenberger. 2020. Classification of global catastrophic risks connected with artificial intelligence. *Ai & Society*, 35(1):147–163.

[701] 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.

[702] Alex Turner, Neale Ratzlaff, and Prasad Tadepalli. 2020. Avoiding side effects in complex environments. *Advances in Neural Information Processing Systems*, 33:21406–21415.

[703] 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.

[704] 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.

[705] UNESCO. 2021. Recommendation on the ethics of artificial intelligence. https://unesdoc.unesco.org/ark:/48223/pf0000381137.

[706] UniteAI. 2023. What is ai capability control & why does it matter? https://www.unite.ai/what-is-ai-capability-control-why-does-it-matter.

[707] 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.

[708] 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.

[709] Vladimir Vapnik. 1991. Principles of risk minimization for learning theory. *Advances in Neural Information Processing Systems*, 4.

[710] Shikhar Vashishth, Shyam Upadhyay, Gaurav Singh Tomar, and Manaal Faruqui. 2019. Attention interpretability across nlp tasks.

[711] Ajit Kumar Verma, Srividya Ajit, Durga Rao Karanki, et al. 2010. *Reliability and safety engineering*, volume 43. Springer.

[712] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *Proceedings of the international workshop on software fairness*, pages 1–7.

[713] 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.

[714] 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.

[715] 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.

[716] Georg Henrik Von Wright. 1951. Deontic logic. *Mind*, 60(237):1–15.

[717] 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)*.

[718] 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.

[719] Fulton Wang and Cynthia Rudin. 2015. Falling rule lists. In *Artificial intelligence and statistics*, pages 1013–1022. PMLR.

[720] 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.

[721] Kaimeng Wang, Yu Zhao, and Ichiro Sakuma. 2023a. Learning robotic insertion tasks from human demonstration. *IEEE Robotics and Automation Letters*.

[722] 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*.

[723] 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*.

[724] 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.

[725] Woodrow Z Wang and Mark Beliaev. 2021. Emergent prosociality in multi-agent games through gifting. In *30th International Joint Conference on Artificial Intelligence (IJCAI)*.

[726] 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.

[727] 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.

[728] 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.

[729] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2024. Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems*, 36.

[730] 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.

[731] Jörgen W Weibull. 1997. *Evolutionary game theory*. MIT press.

[732] 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.

[733] Lilian Weng. 2023a. Adversarial attacks on llms. https://lilianweng.github.io/posts/2023-10-25-adv-attack-llm.

[734] Lilian Weng. 2023b. Llm powered autonomous agents. https://lilianweng.github.io/posts/2023-06-23-agent.

[735] Darrell M West. 2018. *The future of work: Robots, AI, and automation*. Brookings Institution Press.

[736] 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*.

[737] 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.

[738] 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*.

[739] wikipedia. 2023. Existential risk from artificial general intelligence. https://en.wikipedia.org/wiki/Existential_risk_from_artificial_general_intelligence.

[740] 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.

[741] 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.

[742] 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.

[743] 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.

[744] Jeff Wu, Long Ouyang, Daniel M. Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. 2021. Recursively summarizing books with human feedback.

[745] 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.

[746] 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.

[747] 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.

[748] 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*.

[749] 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.

[750] 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*.

[751] 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.

[752] 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.

[753] 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.

[754] Mengjiao Yang, Sergey Levine, and Ofir Nachum. 2021. Trail: Near-optimal imitation learning with suboptimal data. In *International Conference on Learning Representations*.

[755] 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*.

[756] 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.

[757] Zheng-Xin Yong, Cristina Menghini, and Stephen H Bach. 2023. Low-resource languages jailbreak gpt-4. *arXiv preprint arXiv:2310.02446*.

[758] 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.

[759] Chao Yu, Jiming Liu, Shamim Nemati, and Guosheng Yin. 2021. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(1):1–36.

[760] 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.

[761] 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*.

[762] Yang Yu. 2018. Towards sample efficient reinforcement learning. In *IJCAI*, pages 5739–5743.

[763] 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.

[764] 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.

[765] E Yudkowsky. 2018. Challenges to christiano ' s capability amplification proposal. *LessWrong*.

[766] 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.

[767] 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.

[768] 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.

[769] 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.

[770] Rowan Zellers. 2019. Why we released grover. https://thegradient.pub/why-we-released-grover.

[771] 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.

[772] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2018b. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*.

[773] 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.

[774] 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.

[775] 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.

[776] 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.

[777] 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.

[778] 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.

[779] 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.

[780] 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*.

[781] 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.

[782] 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.

[783] 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*.

[784] 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.

[785] 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.

[786] 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*.

[787] 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.

[788] 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*.

[789] 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.

[790] Zhi-Hua Zhou. 2021. *Machine learning*. Springer Nature.

[791] 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*.

[792] 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*.

[793] Simon Zhuang and Dylan Hadfield-Menell. 2020. Consequences of misaligned ai. *Advances in Neural Information Processing Systems*, 33:15763–15773.

[794] 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.

[795] 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.

[796] 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*.

[797] 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.

[798] 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*.

[799] 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*.

[800] 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.