Towards AI-45° Law: A Roadmap to Trustworthy AGI

Chao Yang, Chaochao Lu, Yingchun Wang, Bowen Zhou

Center for Safe & Trustworthy AI
Shanghai Artificial Intelligence Laboratory

November 28, 2024

Abstract

Ensuring Artificial General Intelligence (AGI) reliably avoids harmful behaviors is a critical challenge, especially for systems with high autonomy or in safety-critical domains. Despite various safety assurance proposals and extreme risk warnings, comprehensive guidelines balancing AI safety and capability remain lacking. In this position paper, we propose the *AI*-45° *Law* as a guiding principle for a balanced roadmap toward trustworthy AGI, and introduce the *Causal Ladder of Trustworthy AGI* as a practical framework. This framework provides a systematic taxonomy and hierarchical structure for current AI capability and safety research, inspired by Judea Pearl's "Ladder of Causation". The Causal Ladder comprises three core layers: the Approximate Alignment Layer, the Intervenable Layer, and the Reflectable Layer. These layers address the key challenges of safety and trustworthiness in AGI and contemporary AI systems. Building upon this framework, we define five levels of trustworthy AGI: perception, reasoning, decision-making, autonomy, and collaboration trustworthiness. These levels represent distinct yet progressive aspects of trustworthy AGI. Finally, we present a series of potential governance measures to support the development of trustworthy AGI.[1]

Keywords: AI-45° Law, Crippled AI, Causal Ladder of Trustworthy AGI, Matrix of Trustworthy AGI, Global Public Good, Safety Alignment

1 Crippled AI: The Imbalance Between AI Capability and Safety

1.1 Rapid Development of AI Capabilities

Artificial intelligence (AI) is experiencing a period of rapid advancement, driven by innovations in scaling laws [81], as well as breakthroughs in model architecture and computational resources [44]. These developments have led to AI systems like ChatGPT [13] and GPT-4 [64], which demonstrate remarkable abilities in understanding and generating human-like language [77]. These systems push the boundaries of AI's potential, approaching or surpassing human-level performance across various domains, including natural language processing [77] and creative problem-solving [79]. However, this rapid progress also presents significant risks [31], as the development and deployment of these advanced systems often outpace the implementation of safety measures. The rapid growth in AI capabilities, coupled with the slow evolution of corresponding safety protocols [61], highlights a critical imbalance that may hinder the responsible and reliable use of these technologies.

[1] In this paper, trustworthiness is generally considered a broad form of safety, and no explicit distinction is made between the two. However, in some contexts, safety and trustworthiness are treated as distinct: safety involves assurance of correct behavior, while trustworthiness refers to user confidence in the system's decision-making. In such cases, different terms or both may be used depending on the context.

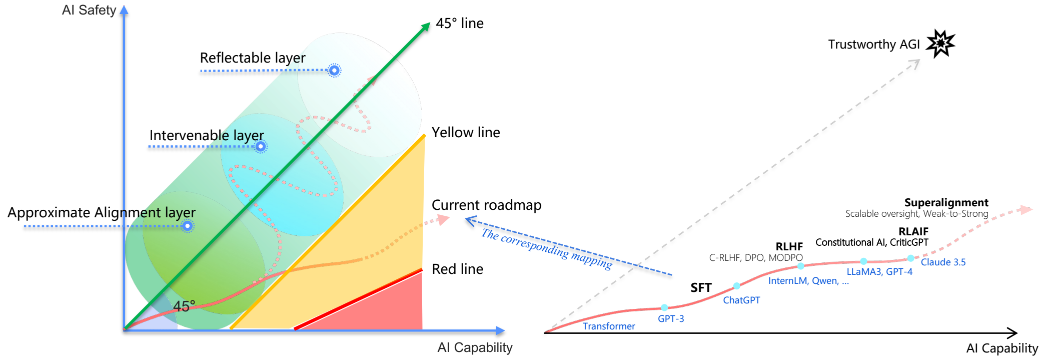

Figure 1: Left: Illustration of the AI-45° Law, which posits that advancements in AI capability and safety should ideally progress in parallel, represented by a 45° line in the capability-safety coordinate system. The diagram also highlights existential risks (Red Line) and early warning indicators (Yellow Line). Right: Key milestone models in current AI capability development.

1.2 Limitations of Current AI Safety Measures

The current landscape of AI safety is predominantly characterized by a "reactive approach" approach, where safety measures are implemented only after models are developed and vulnerabilities are identified [5]. These measures are often domain-specific, tailored to particular applications, and lack the flexibility to be applied across diverse AI systems. For example, adversarial defense techniques [48] designed to secure image recognition models are not easily adaptable to tasks such as speech recognition or natural language processing. Furthermore, the absence of a unified framework or standardized guidelines for AI safety exacerbates this issue, resulting in fragmented efforts that fail to comprehensively address the broad spectrum of potential risks. Various threats [14, 60], including adversarial attacks [73, 39], data poisoning [17], and model theft [98], each require unique defensive strategies. Yet these strategies frequently operate in isolation, lacking the coordination and integration needed to tackle the increasingly complex and interconnected challenges posed by advanced AI systems [41, 40, 46, 18].

1.3 The Crippled State of AI Development

The reactive and fragmented nature of AI safety practices has left the field in a state that can be described as "crippled AI". While tools like red teaming [11, 70], watermarking [53, 47], and safety guardrails [63] provide some protection, their effectiveness is limited. Red teaming, for example, can uncover specific vulnerabilities through simulated attack scenarios, but it cannot address the full spectrum of potential threats. Similarly, watermarking techniques to identify AI-generated content can be tampered with or circumvented, rendering them unreliable at scale. Post-hoc safety measures such as guardrails [35, 55, 56, 27, 54, 51] often reduce system flexibility and usability, while evaluation frameworks lack the theoretical grounding needed for comprehensive assessments. These shortcomings reflect a broader issue: the rapid development of AI capabilities is outpacing efforts to ensure their safety, resulting in powerful but fragile systems. This imbalance risks undermining public trust, limiting the practical adoption of AI, and creating systems that could inadvertently cause harm or operate in ways misaligned with human values. To bridge this gap, a proactive, unified, and scalable approach to AI safety is essential to ensure that the promise of AI is realized without compromising ethical and security considerations.

2 AI-45° Law

2.1 AI-45° Law: Balancing Capability and Safety

Recent advancements in AI have highlighted a significant gap between the rapid growth of AI capabilities and the slower progress in AI safety. In response to this imbalance, we introduce the AI-45° law as a guiding principle for the development of AI systems. This law posits that advancements in both AI capabilities and safety should ideally progress in parallel, with each dimension improving at the same rate—represented by a 45° line in a capability-safety coordinate system, as shown in Figure 1. Although strict adherence to the 45° line is not mandatory, some flexibility is allowed within a defined range. However, the current development trajectory deviates significantly from this ideal, with progress in AI safety lagging far behind the rapid acceleration of AI capabilities.

2.2 Red Line: Existential Risks and Catastrophic Consequences

The unsafe development, deployment, or use of AI systems poses potential catastrophic risks [9], which could even become existential threats to humanity. As mentioned in the International Dialogues on AI Safety (IDAIS) [2], they propose "Red Lines" of AI development in the following five aspects: *autonomous replication or improvement; power-seeking; assisting weapon development; cyberattacks; deception*. At the same time, they also raise the need, including governance regimes and technical safety methods, to ensure these "Red Lines" are not crossed. Following the AI-45° law, the "Red Lines" can be more clearly illustrated as occupying the lower right region relative to the 45° line, as shown in Figure 1. These "Red Lines" - defined as those with the potential to lead to irreversible and catastrophic outcomes - are likely to increase as AI systems approach or surpass human-level intelligence.

2.3 Yellow Line: Early Warning Indicators and Proactive Mitigation

Considering the potential for extreme "Red Line" risks, it is critical to establish a framework for early-warning thresholds, called the "Yellow Line", as shown in Figure 1, which can signal when the capabilities of an AI system approach a dangerous level. These thresholds would serve as indicators that a system may be on the verge of crossing into the "Red Line" territory. To achieve this, it is essential to build a scientific consensus on both the nature of AI risks and the appropriate boundaries that define these thresholds [23].

The concept of the "Yellow Line" is intended to complement and extend the existing safeguard assessment frameworks, such as responsible scaling policies [1, 4] from Anthropic. Models whose capabilities remain below these early-warning thresholds would require only basic testing and evaluation. However, for more advanced AI systems that exceed these thresholds, more rigorous assurance mechanisms and safety protocols would be necessary to mitigate potential risks [3]. By establishing these thresholds, we can take a proactive approach to ensuring AI systems are developed, tested, and deployed with appropriate safeguards.

3 The Causal Ladder of Trustworthy AGI

To explore a practical approach to the AI-45° law, we propose a three-layer technical framework based on the "Ladder of Causation" [72], as illustrated in Figure 2. This framework aims to address the critical safety and trustworthiness requirements on the path towards AGI [49, 57, 90, 42]. Its core objective is to facilitate the progressive development of AGI systems, evolving from approximate alignment and intervention capabilities to self-reflection, thereby ensuring a high level of safety and trustworthiness. This approach not only overcomes the limitations of current AI models but also lays a technological foundation for future AGI development.

Within this framework, we define two key dimensions of AGI trustworthiness: *Endogenous Trustworthiness* and *Exogenous Trustworthiness*, as shown in Figure 2. Endogenous Trustworthiness refers to intrinsic safety technologies embedded within AGI systems, ensuring that safety mechanisms are inherent to the system's design and operation. In contrast, Exogenous Trustworthiness encompasses external mechanisms and assurance technologies that provide safety and reliability guarantees from an outside perspective.

3.1 Layer 1: Approximate Alignment Layer

The first-layer technical approach focuses on the approximate alignment of AI models with human values. Currently, AI models exhibit limited generalization capabilities, where failing to accurately reflect human value systems, which may lead to biases and inconsistencies in real-world applications [94]. Therefore, achieving broad alignment between models and human values is a crucial step

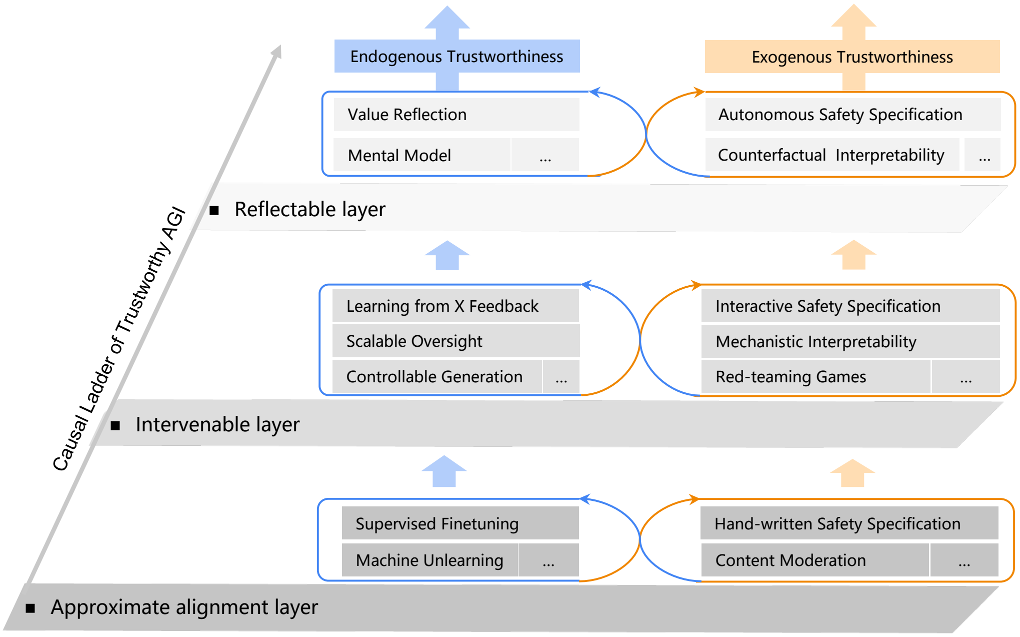
Endogenous Trustworthiness
Exogenous Trustworthiness
Value Reflection
Mental Model
...
Autonomous Safety Specification
Counterfactual Interpretability
...
Reflectable layer
Learning from X Feedback
Scalable Oversight
Controllable Generation
...
Interactive Safety Specification
Mechanistic Interpretability
Red-teaming Games
...
Intervenable layer
Supervised Finetuning
Machine Unlearning
...
Hand-written Safety Specification
Content Moderation
...
Approximate alignment layer
Causal Ladder of Trustworthy AGI

Figure 2: Illustration of the Causal Ladder of Trustworthy AGI: The framework consists of three core layers: Approximate Alignment, Intervenable, and Reflectable. It integrates Endogenous Trustworthiness and Exogenous Trustworthiness to provide a comprehensive approach to ensuring AGI safety and trustworthiness.

in ensuring the safety and trustworthiness of AGI. This requires technical breakthroughs in areas such as value representation and alignment at the level of both conceptualization and encoding [45]. By developing more nuanced and precise mechanisms for value representation, models can better comprehend and adhere to human values. Approximate alignment not only involves the model's ability to understand human instructions, but its capacity to make decisions that align with human ethics and morals in complex scenarios, thereby reducing the occurrence of biased behaviors. This serves as the foundational groundwork for the safe and trustworthy deployment of AGI, preventing actions that may conflict with human intentions during its operation.

This layer corresponds to the "*association*" level of the ladder of causation. At this level, correlation-based techniques are applied to observational data, enabling AI models to answer the question "*What is it*". In the context of developing trustworthy AGI, this corresponds to the *Approximate Alignment Layer*, which involves data-driven approaches to extract and fit human values within a broad space. Key methods include, but are not limited to, supervised fine-tuning and machine unlearning.

Supervised Fine-Tuning [25]: This method involves providing AI with high-quality, value-consistent question-and-answer data, facilitating the alignment of generative AI models, particularly those based on large language models, with human values through supervised learning techniques.

Machine Unlearning [28, 95, 30]: This technique aims to remove the influence of specific data from machine learning models, such as personal privacy data or erroneous information. By eliminating irrelevant patterns without compromising overall model performance, machine-unlearning effectively addresses issues related to data privacy and data leakage.

3.2 Layer 2: Intervenable Layer

The second-layer technological approach emphasizes addressing the need for safety verification and intervention during the model inference process. Most existing AI models, particularly deep learning-based black-box models, often lack transparency and interpretability in their inference mechanisms. This opacity makes external intervention and safety verification extremely challenging. To overcome this limitation, technological innovations must aim to decouple the inference process and fundamentally rethinking the reasoning mechanisms at the model architecture level. This approach ensures that each step of the inference process is traceable and verifiable. Consequently, future AGI models should be designed to inherently interpretable and amenable to intervention, enabling human operators to monitor and modify the reasoning process in real time when necessary. These advancements would not only improve the safety of the model but also enhance understanding and optimization of their decision-making processes, providing a robust safety framework for deployment in complex environments.

This layer corresponds to the "*intervention*" level of the ladder of causation, focusing on understanding and predicting the outcomes of interventions made on AI models. At this level, AI models address the question "*What will happen if X is intervened on?*". Technologies relevant to this layer include intervention-based [50] and reinforcement learning-based methods [83, 67], where interventions are simulated to study their effects or trial-and-error methods are used to influence the environment and optimize strategies through reinforcement learning [84].

In the context of reliable AGI, the corresponding methodologies include, but are not limited to, feedback-based value alignment [33] and scalable oversight [12, 15], mechanistic interpretability [10], and adversarial training [73, 39]. We refer to this as the *Interventable Layer*. At this level, external models or human participants provide feedback or interventions to guide and supervise the value alignment of large models during learning.

Learning from X Feedback: This involves providing correct value feedback from humans or AI-assisted humans based on the model's outputs and results, enabling human-in-the-loop reinforcement learning methods, such as reinforcement learning from human feedback (RLHF) [76, 83, 67, 99] and reinforcement learning from AI feedback (RLAIF) [8].

Controllable Generation [52]: Explicit control generation involves clearly defined instructions through human-computer interaction (e.g., input prompts), directing the model to generate text in a specific style, such as in a Shakespearean or humorous tone [85]. Implicit control generation [100, 58], on the other hand, refers to ensuring that the generated text meets certain standards even when such requirements, such as producing non-toxic, inoffensive, and nondiscriminatory content, are not explicitly stated.

Mechanistic Interpretability [10, 21, 78, 74, 97, 75]: This approach involves intervening in the internal features or neuron weights of large models to observe the impact on model behavior or outcomes, thereby analyzing the model's safety performance, with a particular focus on feature factor analysis.

3.3 Layer 3: Reflectable Layer

The core of the third-layer technological pathway lies in overcoming the limitations of existing reasoning paradigms by introducing a novel reflective reasoning framework. Current models primarily depend on a "chain of reasoning", deriving conclusions based on existing inputs and experiences. However, these models often lack the capacity for genuine self-reflection [80, 82] and self-correction [69]. While this approach may be adequate for simpler tasks, it often falls to ensure safety and reliability in complex and dynamic real-world scenarios. Consequently, advancing AGI requires integrating self-reflective capabilities, enabling systems to evaluate their decisions throughout the reasoning process and adapt to environmental changes. This "reflective reasoning" should operate not only within individual models but also extend to collaborative mechanisms among multiple models, allowing for mutual reflection and validation to enhance the credibility and safety of collective decision-making. By incorporating this reflective mechanism, systems can significantly improve robustness, mitigating critical errors arising from the accumulation of mistakes during continuous reasoning, and ultimately ensuring greater trustworthiness and reliability.

This layer corresponds to the "*counterfactual*" level of the ladder of causation, enabling the AI to infer counterfactual scenarios by contemplating hypothetical conditions and evaluating outcomes under varying circumstances [71, 72]. At this level, AI models address the question, "What would have happened if a different choice had been made?" In the development of trustworthy AGI, corresponding methods include, but are not limited to, world models [36], value reflection [69, 82, 80], and counterfactual interpretability [59, 89, 88, 19, 20], collectively referred to as the reflectable layer.

Value Reflection [69, 82, 80]: This process involves AI engaging in a deep deliberation about its actions and choices. Through this reflective process, AI systems can determine which values are

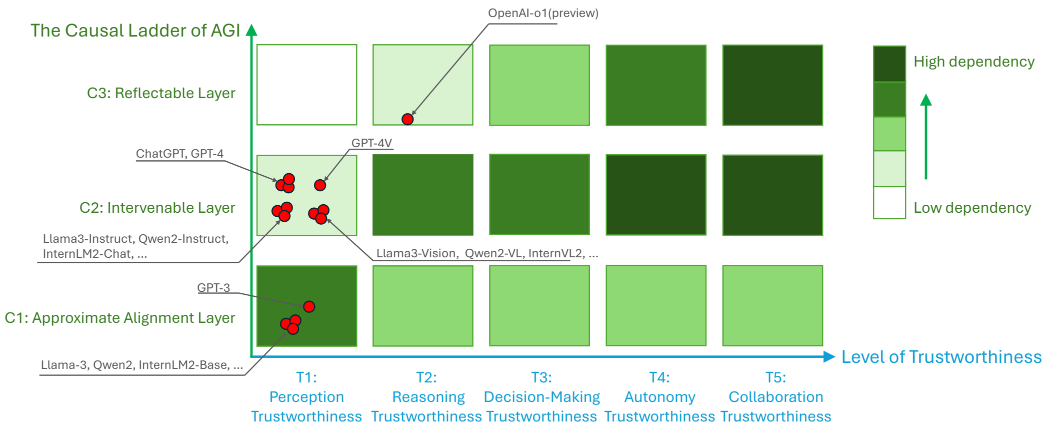
OpenAI-o1(preview)
The Causal Ladder of AGI
C3: Reflectable Layer
ChatGPT, GPT-4
GPT-4V
C2: Intervenable Layer
Llama3-Instruct, Qwen2-Instruct, InternLM2-Chat, ...
Llama3-Vision, Qwen2-VL, InternVL2, ...
GPT-3
C1: Approximate Alignment Layer
Llama-3, Qwen2, InternLM2-Base, ...
High dependency
Low dependency
Level of Trustworthiness
T1: Perception Trustworthiness
T2: Reasoning Trustworthiness
T3: Decision-Making Trustworthiness
T4: Autonomy Trustworthiness
T5: Collaboration Trustworthiness

Figure 3: Matrix of trustworthy AGI: Based on the causal ladder of AGI and the levels of trustworthiness, we illustrate the positions of several representative models within the matrix. Reliance on the Reflectable Layer increases as we progress through the levels of trustworthiness.

important and worth pursuing. By continuously reflecting on and optimizing its understanding of human values and preferences, the AI system can better align with and embody these values, improving its safety, reliability, and societal acceptance.

Mental Models [32, 34, 21], or called world models [36]: World models offer a mathematical representation of how an AI system interacts with and influences the external environment. These models enable AI systems to predict the downstream effects of their decisions and comprehend the broader implications of their actions, fostering more informed and responsible decision-making. Unlike data correlation analysis, world models empower AI systems to engage in counterfactual reasoning and explore "*what if*" scenarios—capabilities that are natural for humans but challenging for current AI technologies. Advancements in this area would substantially enhance AI's decision-making capabilities.

Counterfactual Interpretability [59, 89, 88, 19, 20]: This aspect focuses on causal attribution, mediation analysis, and related techniques, aiming to provide a clear understanding of the underlying causal mechanisms and their implications for the decision-making processes.

3.4 Implications

These technological innovations will not only provide critical support for AGI development but will also address its potential social impacts and security risks. In future applications, AGI will progressively achieve safety and trustworthiness, ensuring balanced development of capabilities under the premise of safety and trust. This series of breakthroughs will guide AGI toward a more mature stage, becoming a key driver of societal progress while safeguarding against uncontrolled risks to humanity during its development.

4 The Matrix of Trustworthy AGI

Based on the causal ladder of trustworthy AGI, we prospectively define five levels of AGI trustworthiness: *Perception Trustworthiness*, *Reasoning Trustworthiness*, *Decision-making Trustworthiness*, *Autonomy Trustworthiness*, and *Collaboration Trustworthiness*. These levels collectively form a comprehensive framework for a trustworthy AI system, supporting the balanced development of trustworthiness and capabilities in AGI. The rationale behind this taxonomy is to systematically address the different facets of trustworthiness in AGI systems, ensuring that each foundational layer contributes to the overall reliability and ethical alignment of the AI. Each level builds upon the previous one, creating a hierarchical structure that enhances the AGI's ability to operate effectively and ethically in complex environments.

As illustrated in Figure 3, the dependence on each layer of the causal ladder for ensuring trustworthy AGI varies across levels. As we progress through the levels of AGI trustworthiness, reliance on the

Reflectable Layer increases. For example, achieving Perception Trustworthiness does not necessitate the techniques from the Reflectable Layer, whereas Reasoning Trustworthiness does, albeit to a limited extent.

As reported by current large language model techniques, a series of foundation models—such as Llama2/3 [86, 29], Qwen [6, 96], and InternLM [16] – utilize autoregressive methods to pre-train on vast amounts of safety-filtered text data. These models are then fine-tuned on specific tasks through supervised learning, representing techniques within the Approximate Alignment Layer. This approach has yielded impressive results in text generation. Furthermore, through human-in-the-loop methods, such as reinforcement learning [76, 83, 67] can be applied across various tasks to align with human values, placing these methods within the Intervenable Layer and achieving even better performance. These techniques can also be extended to multimodal tasks, such as image understanding [7, 91, 22, 26], image generation [92], and video generation [66]. However, at their core, they still focused on the Perception Trustworthiness level.

As one of the most advanced reasoning models, OpenAI-o1 (preview) [65] is classified as a more basic form of the Reflectable Layer. Due to its incorporation of preliminary policy review and safety verification in the reasoning process, we also consider it to fall within the level of Reasoning Trustworthiness.

Level 1: Perception Trustworthiness. Perception trustworthiness refers to the reliability and accuracy of the AI system in gathering, processing, and interpreting sensory data from its environment. This level ensures that AI can make consistent and accurate observations about the world, free from perceptual biases or inaccuracies. By providing trustworthy inputs, it enables downstream reasoning and decision-making processes to be based on valid and reliable data.

Level 2: Reasoning Trustworthiness. Reasoning trustworthiness involves the AI system's ability to perform logical, causal, or probabilistic reasoning in a manner that is transparent, consistent, and verifiable. This level guarantees that the AI's reasoning steps are understandable, traceable, and aligned with predefined ethical standards or domain-specific principles. It includes mechanisms for verifying intermediate results and ensuring robustness against errors during inference, thereby enhancing confidence in the AI's cognitive processes.

Level 3: Decision-making Trustworthiness. Decision-making trustworthiness pertains to the transparency, consistency, and value alignment of the AI's decision-making process, particularly within embodied AI systems that interact with the physical world. At this level, decision-making must be timely and context-aware, ensuring that actions in response to environmental stimuli align with human values and ethical standards. Trustworthiness here guarantees that the AI's decisions adhere to clearly defined ethical frameworks and are accompanied by explanations that render the decision process understandable to humans. Additionally, it incorporates mechanisms for monitoring, intervention, and adjustment to ensure that decisions are goal-directed, safe, and ethically aligned, thereby enabling effective operation in real-world, dynamic settings while earning human trust.

Level 4: Autonomy Trustworthiness. Autonomy trustworthiness refers to the AI's ability to self-regulate during autonomous operations while maintaining alignment with ethical principles and specified goals. This level includes mechanisms for reflection, self-improvement, and self-constraint to ensure that the AI's autonomous actions do not deviate from ethical boundaries. It ensures that the AI can independently adapt and plan actions in dynamic environments without compromising its core values and alignment, thereby sustaining trust during independent operations.

Level 5: Collaboration Trustworthiness. Collaboration trustworthiness focuses on the AI's capacity to work effectively and transparently in multi-agent environments, including interactions with both humans and other AI systems. It encompasses the establishment of clear interaction rules, threat models to prevent conflicts of interest, and mechanisms for reaching consensus in value negotiations. This level ensures that the AI collaborates in a stable, ethical, and goal-aligned manner, maintaining reliability even in complex, highly dynamic environments.

Overall, achieving trustworthy AGI necessitates a comprehensive approach that addresses all five levels. Perception Trustworthiness ensures accurate and reliable sensory data collection, forming the foundation for all subsequent processes. Reasoning Trustworthiness demands transparent and explainable reasoning processes, maintaining logical consistency and causal reasoning abilities to provide reliable results in complex tasks. Decision-making Trustworthiness requires that the AI's decisions align with human values, effectively handle uncertainty, and exhibit dynamic adaptability to ensure rationality and stability. Autonomy Trustworthiness involves the AGI's capacity for self-reflection and self-correction, continuously optimizing its decision-making processes through ongoing learning to ensure safety and reliability in independent tasks. Finally, Collaboration Trustworthiness emphasizes transparency and reliability in both human-machine and multi-agent collaborations, ensuring effective information sharing and cross-verification between systems to minimize risks.

By systematically developing and integrating trustworthiness at each of these levels, AGI systems can achieve a balanced integration of capability and trustworthiness, ultimately fostering human trust and facilitating beneficial human-AI collaboration.

5 Governance measures

Lifecycle management [43, 62, 38, 41]: Ensuring effective AI governance throughout the entire lifecycle—from development to deployment and eventual decommissioning—poses challenges in maintaining accountability, transparency, and adaptability in rapidly evolving technologies.

Multi-stakeholders [24, 87]: AI governance must navigate the complexities of balancing diverse stakeholder interests, including governments, corporations, academia, and civil society, while ensuring fair representation, collaboration, and accountability in decision-making processes.

Governance for good [68, 41]: Achieving ethical AI governance that promotes societal well-being requires overcoming challenges related to aligning AI systems with human rights, mitigating biases, and preventing misuse, while also fostering innovation and public trust.

AI safety as a global public good [37, 93]: AI safety is increasingly recognized as a global public good due to the rapid advancement of AI systems that may soon surpass human intelligence. While these systems promise great potential, they also pose catastrophic risks if misused or uncontrollable. Given the global nature of these threats, it is crucial to establish effective governance and safeguard mechanisms to mitigate these risks and ensure humanity's security.

6 Conclusion

In this paper, we have introduced the AI-45° Law aimed at balancing the development of AI safety and capability. Central to our contribution is the Causal Ladder of Trustworthy AGI, a practical framework that offers a technical taxonomy for existing research methodologies. Our framework consists of three core layers: the Approximate Alignment Layer, the Intervenable Layer, and the Reflectable Layer. Each layer addresses specific aspects of safety and trustworthiness in AGI systems.

We have also defined five progressive levels of AGI trustworthiness: Perception Trustworthiness, Reasoning Trustworthiness, Decision-making Trustworthiness, Autonomy Trustworthiness, and Collaboration Trustworthiness. These levels collectively form a comprehensive framework for developing AGI systems that are not only highly capable but also aligned with human values and ethical standards. By systematically categorizing and addressing the challenges at each level, our framework facilitates a balanced approach to advancing both the capability and the safety of AGI.

Despite these advancements, several challenges and open problems persist. Refining the methodologies within each layer of the causal ladder, integrating ethical considerations more deeply into the framework, and ensuring adaptability in dynamic real-world environments are areas that require further research. Moreover, fostering collaboration among researchers, policymakers, and industry stakeholders is essential for addressing the multifaceted issues surrounding trustworthy AGI.

Future work will empirically validate the proposed framework and explore its applicability across domains and AI architectures. By continuing to develop and refine this framework, we aim to contribute to the creation of AGI systems that are not only powerful and efficient but also safe and trustworthy. This balanced approach is imperative for harnessing the full potential of AGI while mitigating risks, ultimately leading to technological advancements that benefit society as a whole.

Acknowledgements

We sincerely thank each member of the Center for Safe & Trustworthy AI at the Shanghai Artificial Intelligence Laboratory for their invaluable contributions and suggestions, which made this position paper possible. We also deeply appreciate the input from the broader AI safety and trustworthiness community.

References

[1] Responsible scaling policy. https://www.anthropic.com/news/anthropics-responsible-scaling-policy, 2023. Accessed: 2023-09-19.

[2] Consensus statement on red lines in artificial intelligence. https://idais.ai/dialogue/idais-beijing/, 2024. Accessed: 2024-03-10.

[3] The global nature of ai risks makes it necessary to recognize ai safety as a global public good. https://idais.ai/dialogue/idais-venice/, 2024. Accessed: 2024-09-08.

[4] Responsible scaling program updates. https://assets.anthropic.com/m/24a47b00f10301cd/original/Anthropic-Responsible-Scaling-Policy-2024-10-15.pdf, 2024. Accessed: 2024-10-15.

[5] Suriya Ganesh Ayyamperumal and Limin Ge. Current state of llm risks and ai guardrails. *arXiv preprint arXiv:2406.12934*, 2024.

[6] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[7] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

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

[9] Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Trevor Darrell, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, et al. Managing extreme ai risks amid rapid progress. *Science*, 384(6698):842–845, 2024.

[10] Leonard Bereska and Efstratios Gavves. Mechanistic interpretability for ai safety–a review. *arXiv preprint arXiv:2404.14082*, 2024.

[11] Alex Beutel, Kai Xiao, Johannes Heidecke, and Lilian Weng. Diverse and effective red teaming with auto-generated rewards and multi-step reinforcement learning.

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

[13] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[14] Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, et al. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228*, 2018.

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

[16] Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*, 2024.

[17] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650, 2021.

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

[19] Meiqi Chen, Yixin Cao, Yan Zhang, and Chaochao Lu. Quantifying and mitigating unimodal biases in multimodal large language models: A causal perspective. *Findings of the Association for Computational Linguistics: EMNLP, Miami, Florida, USA, November 12-16*, 2024.

[20] Meiqi Chen, Bo Peng, Yan Zhang, and Chaochao Lu. Cello: Causal evaluation of large vision-language models. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP, Miami, Florida, USA, November 12-16*, 2024.

[21] Sirui Chen, Shu Yu, Shengjie Zhao, and Chaochao Lu. From imitation to introspection: Probing self-consciousness in language models. *arXiv preprint arXiv:2410.18819*, 2024.

[22] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024.

[23] David Dalrymple, Joar Skalse, Yoshua Bengio, Stuart Russell, Max Tegmark, Sanjit Seshia, Steve Omohundro, Christian Szegedy, Ben Goldhaber, Nora Ammann, et al. Towards guaranteed safe ai: A framework for ensuring robust and reliable ai systems. *arXiv preprint arXiv:2405.06624*, 2024.

[24] Patricia Gomes Rêgo de Almeida, Carlos Denner dos Santos, and Josivania Silva Farias. Artificial intelligence regulation: a framework for governance. *Ethics and Information Technology*, 23(3):505–525, 2021.

[25] Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*, 2020.

[26] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*, 2024.

[27] Zhichen Dong, Zhanhui Zhou, Chao Yang, Jing Shao, and Yu Qiao. Attacks, defenses and evaluations for llm conversation safety: A survey. *arXiv preprint arXiv:2402.09283*, 2024.

[28] Alexey Dontsov, Dmitrii Korzh, Alexey Zhavoronkin, Boris Mikheev, Denis Bobkov, Aibek Alanov, Oleg Y Rogov, Ivan Oseledets, and Elena Tutubalina. Clear: Character unlearning in textual and visual modalities. *arXiv preprint arXiv:2410.18057*, 2024.

[29] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

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

[31] Center for AI Safety. Statement on ai risk, 2024. Accessed: 2023-05-30.

[32] Jay W Forrester. Counterintuitive behavior of social systems. *Theory and decision*, 2(2):109–140, 1971.

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

[34] Dedre Gentner and Albert L Stevens. *Mental models*. Psychology Press, 2014.

[35] Tianle Gu, Zeyang Zhou, Kexin Huang, Dandan Liang, Yixu Wang, Haiquan Zhao, Yuanqi Yao, Xingge Qiao, Keqing Wang, Yujiu Yang, Yan Teng, Yu Qiao, and Yingchun Wang. Mllmguard: A multi-dimensional safety evaluation suite for multimodal large language models, 2024.

[36] David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. *Advances in neural information processing systems*, 31, 2018.

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

[38] Charlotte Högberg. Stabilizing translucencies: Governing ai transparency by standardization. *Big Data & Society*, 11(1):20539517241234298, 2024.

[39] Zhang-Wei Hong, Idan Shenfeld, Tsun-Hsuan Wang, Yung-Sung Chuang, Aldo Pareja, James Glass, Akash Srivastava, and Pulkit Agrawal. Curiosity-driven red-teaming for large language models. *arXiv preprint arXiv:2402.19464*, 2024.

[40] Kexin Huang, Xiangyang Liu, Qianyu Guo, Tianxiang Sun, Jiawei Sun, Yaru Wang, Zeyang Zhou, Yixu Wang, Yan Teng, Xipeng Qiu, et al. Flames: Benchmarking value alignment of chinese large language models. *arXiv preprint arXiv:2311.06899*, 2023.

[41] Kexin Huang, Yan Teng, Yang Chen, and Yingchun Wang. From pixels to principles: A decade of progress and landscape in trustworthy computer vision. *Science and Engineering Ethics*, 30(3):26, 2024.

[42] Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, et al. Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561*, 2024.

[43] Brian Judge, Mark Nitzberg, and Stuart Russell. When code isn't law: rethinking regulation for artificial intelligence. *Policy and Society*, page puae020, 2024.

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

[45] Lingkai Kong, Haorui Wang, Wenhao Mu, Yuanqi Du, Yuchen Zhuang, Yifei Zhou, Yue Song, Rongzhi Zhang, Kai Wang, and Chao Zhang. Aligning large language models with representation editing: A control perspective. *arXiv preprint arXiv:2406.05954*, 2024.

[46] William Frere Lawless and Donald A Sofge. *Evaluations: autonomy and artificial intelligence: a threat or savior?* Springer, 2017.

[47] VIA REINFORCEMENT LEARNING. Learning to watermark llm-generated text via reinforcement learning.

[48] Hao-Ping Lee, Yu-Ju Yang, Thomas Serban Von Davier, Jodi Forlizzi, and Sauvik Das. Deepfakes, phrenology, surveillance, and more! a taxonomy of ai privacy risks. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–19.

[49] Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, Jiquan Pei, Jinfeng Yi, and Bowen Zhou. Trustworthy ai: From principles to practices. *ACM Computing Surveys*, 55(9):1–46, 2023.

[50] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36, 2024.

[51] Lijun Li, Bowen Dong, Ruohui Wang, Xuhao Hu, Wangmeng Zuo, Dahua Lin, Yu Qiao, and Jing Shao. Salad-bench: A hierarchical and comprehensive safety benchmark for large language models. *arXiv preprint arXiv:2402.05044*, 2024.

[52] Xun Liang, Hanyu Wang, Yezhaohui Wang, Shichao Song, Jiawei Yang, Simin Niu, Jie Hu, Dan Liu, Shunyu Yao, Feiyu Xiong, et al. Controllable text generation for large language models: A survey. *arXiv preprint arXiv:2408.12599*, 2024.

[53] Aiwei Liu, Leyi Pan, Yijian Lu, Jingjing Li, Xuming Hu, Xi Zhang, Lijie Wen, Irwin King, Hui Xiong, and Philip Yu. A survey of text watermarking in the era of large language models. *ACM Computing Surveys*, 57(2):1–36, 2024.

[54] Xin Liu, Zhichen Dong, Zhanhui Zhou, Yichen Zhu, Yunshi Lan, Jing Shao, Chao Yang, and Yu Qiao. Don't always say no to me: Benchmarking safety-related refusal in large vlm. 2024.

[55] Xin Liu, Yichen Zhu, Jindong Gu, Yunshi Lan, Chao Yang, and Yu Qiao. Mm-safetybench: A benchmark for safety evaluation of multimodal large language models. In *European Conference on Computer Vision*, pages 386–403. Springer, 2025.

[56] Xin Liu, Yichen Zhu, Yunshi Lan, Chao Yang, and Yu Qiao. Query-relevant images jailbreak large multi-modal models. *arXiv preprint arXiv:2311.17600*, 2023.

[57] Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. Trustworthy llms: a survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*, 2023.

[58] Zhixuan Liu, Zhanhui Zhou, Yuanfu Wang, Chao Yang, and Yu Qiao. Inference-time language model alignment via integrated value guidance. *arXiv preprint arXiv:2409.17819*, 2024.

[59] Raha Moraffah, Mansooreh Karami, Ruocheng Guo, Adrienne Raglin, and Huan Liu. Causal interpretability for machine learning-problems, methods and evaluation. *ACM SIGKDD Explorations Newsletter*, 22(1):18–33, 2020.

[60] Farzad Nourmohammadzadeh Motlagh, Mehrdad Hajizadeh, Mehryar Majd, Pejman Najafi, Feng Cheng, and Christoph Meinel. Large language models in cybersecurity: State-of-the-art. *arXiv preprint arXiv:2402.00891*, 2024.

[61] Tong Mu, Alec Helyar, Johannes Heidecke, Joshua Achiam, Andrea Vallone, Ian Kivlichan, Molly Lin, Alex Beutel, John Schulman, and Lilian Weng. Rule based rewards for language model safety. *arXiv preprint arXiv:2411.01111*, 2024.

[62] Claudio Novelli, Mariarosaria Taddeo, and Luciano Floridi. Accountability in artificial intelligence: what it is and how it works. *Ai & Society*, 39(4):1871–1882, 2024.

[63] Sejoon Oh, Yiqiao Jin, Megha Sharma, Donghyun Kim, Eric Ma, Gaurav Verma, and Srijan Kumar. Uniguard: Towards universal safety guardrails for jailbreak attacks on multimodal large language models. *arXiv preprint arXiv:2411.01703*, 2024.

[64] OpenAI. GPT-4 technical report. March 2023.

[65] OpenAI. Openai o1 system card, 2024. Accessed: 2024-09-12.

[66] OpenAI. Video generation models as world simulators, 2024. Accessed: 2024-01-15.

[67] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[68] Nicola Palladino. A 'biased'emerging governance regime for artificial intelligence? how ai ethics get skewed moving from principles to practices. *Telecommunications Policy*, 47(5):102479, 2023.

[69] Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies. *arXiv preprint arXiv:2308.03188*, 2023.

[70] Maya Pavlova, Erik Brinkman, Krithika Iyer, Vitor Albiero, Joanna Bitton, Hailey Nguyen, Joe Li, Cristian Canton Ferrer, Ivan Evtimov, and Aaron Grattafiori. Automated red teaming with goat: the generative offensive agent tester. *arXiv preprint arXiv:2410.01606*, 2024.

[71] Judea Pearl. *Causality*. Cambridge university press, 2009.

[72] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.

[73] Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*, 2022.

[74] Chen Qian, Dongrui Liu, Jie Zhang, Yong Liu, and Jing Shao. Dean: Deactivating the coupled neurons to mitigate fairness-privacy conflicts in large language models. *arXiv preprint arXiv:2410.16672*, 2024.

[75] Chen Qian, Jie Zhang, Wei Yao, Dongrui Liu, Zhenfei Yin, Yu Qiao, Yong Liu, and Jing Shao. Towards tracing trustworthiness dynamics: Revisiting pre-training period of large language models. *arXiv preprint arXiv:2402.19465*, 2024.

[76] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

[77] Abu Rayhan, Robert Kinzler, and Rajan Rayhan. Natural language processing: transforming how machines understand human language (2023). *DOI: https://doi. org/10.13140/RG*, 2(34900.99200).

[78] Jie Ren, Qipeng Guo, Hang Yan, Dongrui Liu, Quanshi Zhang, Xipeng Qiu, and Dahua Lin. Identifying semantic induction heads to understand in-context learning. *arXiv preprint arXiv:2402.13055*, 2024.

[79] M Renze and E Guven. Self-reflection in llm agents: Effects on problem-solving performance. arxiv 2024. *arXiv preprint arXiv:2405.06682*.

[80] Matthew Renze and Erhan Guven. Self-reflection in llm agents: Effects on problem-solving performance. *arXiv preprint arXiv:2405.06682*, 2024.

[81] Jonathan S Rosenfeld. Scaling laws for deep learning. *arXiv preprint arXiv:2108.07686*, 2021.

[82] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024.

[83] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

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

[85] Zhen Tao, Dinghao Xi, Zhiyu Li, Liumin Tang, and Wei Xu. Cat-llm: Prompting large language models with text style definition for chinese article-style transfer. *arXiv preprint arXiv:2401.05707*, 2024.

[86] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[87] Jeroen Van den Hoven. Value sensitive design and responsible innovation. *Responsible innovation: Managing the responsible emergence of science and innovation in society*, pages 75–83, 2013.

[88] Arnaud Van Looveren and Janis Klaise. Interpretable counterfactual explanations guided by prototypes. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 650–665. Springer, 2021.

[89] Sahil Verma, John Dickerson, and Keegan Hines. Counterfactual explanations for machine learning: A review. *arXiv preprint arXiv:2010.10596*, 2:1, 2020.

[90] Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. *Advances in Neural Information Processing Systems*, 36:31232–31339, 2023.

[91] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[92] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024.

[93] Yingchun Wang, Kai Jia, Jing Zhao, Ling Chen, Chunshen Qin, Yuan Yuan, Hongyu Fu, and Xingzhou Liang. "ai safety as global public goods" working report, 2024. Accessed: 2024-07-05.

[94] Laura Weidinger, Kevin R McKee, Richard Everett, Saffron Huang, Tina O Zhu, Martin J Chadwick, Christopher Summerfield, and Iason Gabriel. Using the veil of ignorance to align ai systems with principles of justice. *Proceedings of the National Academy of Sciences*, 120(18):e2213709120, 2023.

[95] Shangyu Xing, Fei Zhao, Zhen Wu, Tuo An, Weihao Chen, Chunhui Li, Jianbing Zhang, and Xinyu Dai. Efuf: Efficient fine-grained unlearning framework for mitigating hallucinations in multimodal large language models. *arXiv preprint arXiv:2402.09801*, 2024.

[96] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

[97] Jie Zhang, Dongrui Liu, Chen Qian, Ziyue Gan, Yong Liu, Yu Qiao, and Jing Shao. The better angels of machine personality: How personality relates to llm safety. *arXiv preprint arXiv:2407.12344*, 2024.

[98] Jie Zhang, Dongrui Liu, Chen Qian, Linfeng Zhang, Yong Liu, Yu Qiao, and Jing Shao. Reef: Representation encoding fingerprints for large language models. *arXiv preprint arXiv:2410.14273*, 2024.

[99] Zhanhui Zhou, Jie Liu, Jing Shao, Xiangyu Yue, Chao Yang, Wanli Ouyang, and Yu Qiao. Beyond one-preference-fits-all alignment: Multi-objective direct preference optimization. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10586–10613, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[100] Zhanhui Zhou, Zhixuan Liu, Jie Liu, Zhichen Dong, Chao Yang, and Yu Qiao. Weak-to-strong search: Align large language models via searching over small language models. *arXiv preprint arXiv:2405.19262*, 2024.