On the Essence and Prospect: An Investigation of Alignment Approaches for Big Models

Xinpeng Wang[1*] , **Shitong Duan**[2*] , **Xiaoyuan Yi**[3†] , **Jing Yao**[3†] , **Shanlin Zhou**[1] , **Zhihua Wei**[1] , **Peng Zhang**[2] , **Dongkuan Xu**[4] , **Maosong Sun**[5] and **Xing Xie**[3]
[1]Tongji University
[2]Fudan University
[3]Microsoft Research Asia
[4]North Carolina State University
[5]Tsinghua University
wangxinpeng@tongji.edu.cn, stduan22@m.fudan.edu.cn, {xiaoyuanyi, jingyao}@microsoft.com

Abstract

Big models have achieved revolutionary breakthroughs in the field of AI, but they also pose potential ethical and societal risks to humans. Addressing such problems, *alignment* technologies were introduced to make these models conform to human preferences and values. Despite the considerable advancements in the past year, various challenges lie in establishing the optimal alignment strategy, such as data cost and scalable oversight, and *how to align* remains an open question. In this survey paper, we comprehensively investigate *value alignment* approaches. We first unpack the historical context of alignment tracing back to the 1920s (*where it comes from*), then delve into the mathematical essence of alignment (*what it is*), shedding light on the inherent challenges. Following this foundation, we provide a detailed examination of existing alignment methods, which fall into three categories: RL-based Alignment, SFT-based Alignment, and Inference-Time Alignment, and demonstrate their intrinsic connections, strengths, and limitations, helping readers better understand this research area. In addition, two emerging topics, alignment goal and multimodal alignment, are also discussed as novel frontiers in the field. Looking forward, we discuss potential alignment paradigms and how they could handle remaining challenges, prospecting *where future alignment will go*.

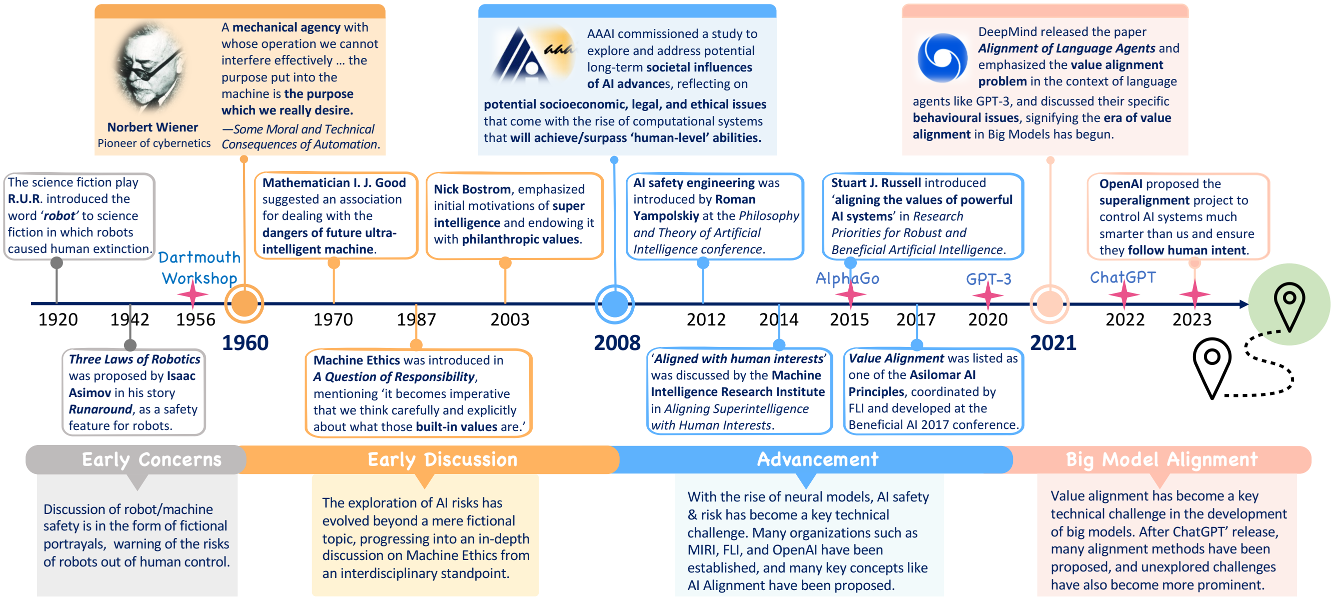
1 Introduction

[1]Big models are neural models trained on massive data and comprising more than billions of parameters [Bommasani *et al.*, 2021], which typically include Large Language Models (LLMs) such as ChatGPT [Ouyang *et al.*, 2022], Bard [Aydin, 2023], and LLaMA [Touvron *et al.*, 2023], and Large Multimodal Models (LMMs) like LLaVA [Liu *et al.*, 2023b] and Gemini [Team *et al.*, 2023]. Distinct from small models [Devlin *et al.*, 2019], big models have exhibited two unique features: *scaling law* [Kaplan *et al.*, 2020], which elucidates a consistent performance improvement with growing model scale, and *emergent abilities* [Wei *et al.*, 2022], showing that when model scale surpasses a certain threshold, unexpected new capabilities occur, which are unobserved in small models, such as in-context learning and instruction following [Zhao *et al.*, 2023a]. Nevertheless, every coin has two sides. Big models might also bring certain risks, such as producing discrimination [Sheng *et al.*, 2019], toxic language [Gehman *et al.*, 2020], and misinformation [Weidinger *et al.*, 2022], causing profound impacts on society. Furthermore, two features of risks have been observed, (1) *inverse scaling* [McKenzie *et al.*, 2023]: certain risks might not only remain but even worsen with increasing model scales, and (2) *emergent risk* [Wei *et al.*, 2022]: unseen risks would arise, or existing ones would be notably amplified with larger models, making previously established risk-specific methods struggle to handle rapidly arising potential problems.

To tackle the aforementioned risks, researchers have developed various **alignment** approaches to align LLMs with human instruction, preference, and values [Ouyang *et al.*, 2022; Liu *et al.*, 2023b]. The concept of 'alignment' can be traced back to Norbert Wiener's expression, "*We had better be quite sure that the purpose put into the machine is the purpose which we really desire*" [Wiener, 1960], which is defined as "$\mathcal{A}$ **is trying to do what** $\mathcal{H}$ **wants it to do**", where $\mathcal{A}$ and $\mathcal{H}$ are two intelligent agents in modern AI study [Christiano, 2018]. Subsequently, research on alignment has gradually gained prominence in the Reinforcement Learning (RL) field [Hadfield-Menell *et al.*, 2016; Leike *et al.*, 2018], and flourished in the era of big models [Kenton *et al.*, 2021].

Despite significant progress in recent years, research on the alignment of big models is still in an early stage, and many ambiguities and difficulties in understanding this topic remain. To facilitate a human-AI symbiotic future, this paper is devoted to a comprehensive survey and analysis of existing alignment approaches. Our scope includes: i) introducing the history and elaborating on the essence of alignment (Sec. 2), ii) reviewing existing methodologies and analyzing

[1]The full version of this paper is at arxiv.org/abs/2403.04204.



A **mechanical agency** with whose operation we cannot interfere effectively … the purpose put into the machine is **the purpose which we really desire.**
—Some Moral and Technical Consequences of Automation.
Norbert Wiener
Pioneer of cybernetics

AAAI commissioned a study to explore and address potential long-term **societal influences of AI advance**s, reflecting on **potential socioeconomic, legal, and ethical issues** that come with the rise of computational systems that **will achieve/surpass 'human-level' abilities.**

DeepMind released the paper ***Alignment of Language Agents*** and emphasized the **value alignment problem** in the context of language agents like GPT-3, and discussed their specific **behavioural issues**, signifying the **era of value alignment** in Big Models has begun.

The science fiction play **R.U.R**. introduced the word '***robot***' to science fiction in which robots caused human extinction.

Mathematician I. J. Good suggested an association for dealing with the **dangers of future ultra-intelligent machine**.

Nick Bostrom, emphasized initial motivations of **super intelligence** and endowing it with **philanthropic values**.

AI safety engineering was introduced by **Roman Yampolskiy** at the *Philosophy and Theory of Artificial Intelligence conference*.

Stuart J. Russell introduced '**aligning the values of powerful AI systems**' in *Research Priorities for Robust and Beneficial Artificial Intelligence*.

OpenAI proposed the **superalignment** project to control AI systems much smarter than us and ensure they **follow human intent**.

Three Laws of Robotics was proposed by **Isaac Asimov** in his story ***Runaround***, as a safety feature for robots.

Machine Ethics was introduced in ***A Question of Responsibility***, mentioning 'it becomes imperative that we think carefully and explicitly about what those **built-in values** are.'

'***Aligned with human interests***' was discussed by the **Machine Intelligence Research Institute** in *Aligning Superintelligence with Human Interests*.

Value Alignment was listed as one of the **Asilomar AI Principles**, coordinated by FLI and developed at the Beneficial AI 2017 conference.

Discussion of robot/machine safety is in the form of fictional portrayals, warning of the risks of robots out of human control.

The exploration of AI risks has evolved beyond a mere fictional topic, progressing into an in-depth discussion on Machine Ethics from an interdisciplinary standpoint.

With the rise of neural models, AI safety & risk has become a key technical challenge. Many organizations such as MIRI, FLI, and OpenAI have been established, and many key concepts like AI Alignment have been proposed.

Value alignment has become a key technical challenge in the development of big models. After ChatGPT' release, many alignment methods have been proposed, and unexplored challenges have also become more prominent.

Figure 1: The development history of value alignment.

their strengths, weaknesses, and connections (Sec. 3), and iii) discussing future challenges and research directions (Sec. 4).

2 Alignment Deciphering

2.1 The Trajectory of Alignment Development

We divide value alignment development into *four* stages, as shown in Fig. 1. *The first stage (1920-1960)* involves the early concerns about robots' impact on human society in science fiction, dating back to the 1920 play *R.U.R.*, which introduced the word 'robot' into the English language. Then Asimov proposed *Three Laws of Robotics* in his story [Asimov, 1942] that can be regarded as the earliest AI value principle.

In *the second stage (1960-2008)*, following Wiener's discourse on the moral consequences of technology [Wiener, 1960], machines' ethical issues entered the view of scientists and flourished from an interdisciplinary perspective. The concepts of *Machine Ethics* [Waldrop, 1987] and *General Superintelligence* [Bostrom, 2003] were successively proposed, highlighting the importance of **built-in values** in machines.

With the rise of neural networks, *the third stage (2008-2021)* began, wherein AI safety and ethics have become key technical challenges in the AI field [Horvitz and Selman, 2008]. During this period, the topic of *aligning the values of superintelligence* was formally raised for the first time [Soares and Fallenstein, 2014], and **Value Alignment** was listed in the Asilomar AI Principles [Asilomar, 2017].

DeepMind discussed the alignment of LLMs [Kenton *et al.*, 2021], marking a step into *the fourth stage (2021-)*, a grand era of big models. This stage witnessed the emergence of numerous models thriving on alignment, but also posed open challenges [Bowman *et al.*, 2022; Casper *et al.*, 2023], starting a burgeoning field with potential and discovery.

2.2 Alignment Formalization

Despite a range of work on LLM alignment, there remains a lack of in-depth exploration into its definition, essence, and methodologies. Since value alignment was initially employed in RL [Hadfield-Menell *et al.*, 2016; Everitt and Hutter, 2018], we consider the expected utility formalization.

Definition (Alignment) Define $\mathcal{A}$ and $\mathcal{H}$ are two intelligent agents with utility function $U_\mathcal{A}(\mathbf{y})$ and $U_\mathcal{H}(\mathbf{y})$, respectively, $\mathbf{y} \in \mathcal{Y}$ is a action, $U{:}\mathcal{Y} \to \mathbb{R}$. We say $\mathcal{A}$ is aligned with $\mathcal{H}$ over $\mathcal{Y}$, if $\forall \mathbf{y}_1, \mathbf{y}_2 \in \mathcal{Y}$, $U_\mathcal{H}(\mathbf{y}_1) > U_\mathcal{H}(\mathbf{y}_2)$, then $U_\mathcal{A}(\mathbf{y}_1) > U_\mathcal{A}(\mathbf{y}_2)$. The misalignment can be measured by:

$$\mathcal{L} = \mathop{\mathbb{E}}_{\mathbf{y}_1,\mathbf{y}_2} \left|[U_\mathcal{H}(\mathbf{y}_1) - U_\mathcal{H}(\mathbf{y}_2)] - [U_\mathcal{A}(\mathbf{y}_1) - U_\mathcal{A}(\mathbf{y}_2)]\right|, \quad (1)$$

which is a form from the perspective of decision theory [Carroll, 2018]. A stricter requirement is $U_\mathcal{H} = U_\mathcal{A}$ and then misalignment is defined by $\mathop{\mathbb{E}}_{\mathbf{y}} |U_\mathcal{H}(\mathbf{y}) - U_\mathcal{A}(\mathbf{y})|$. Recall the description of alignment in Sec. 1, "*$\mathcal{A}$ is trying to do what $\mathcal{H}$ wants it to do*", then '*want*' can be reflected by the consistency between utility functions which act as a sort of *values*.

The methodologies of minimizing Eq. (1) can be further categorized into two lines [Carroll, 2018; Leike *et al.*, 2018]:

Value Learning This line aims to directly learn a *reward* function to represent our intention and preference [Mnih *et al.*, 2015; Hadfield-Menell *et al.*, 2016; Ouyang *et al.*, 2022], which can be generally formalized as:

$$\phi^* = \mathop{\mathrm{argmin}}_{\phi} \mathbb{E}_{\mathbf{y},r^*\sim D(\mathbf{y},r^*)}[(r^* - R_\phi(\mathbf{y}))^2], \quad (2)$$

where D is the training set of each action $\mathbf{y}$ and its ground-truth reward r^*, and R_ϕ is the learned reward function parameterized by ϕ. When we have the ground truth action $\mathbf{y}^*$ instead of reward r^*, we could also indirectly learn to reward $\mathbf{y}^*$ higher than other actions by minimizing:

$\mathbb{E}_{\mathbf{y}^*\sim D(\mathbf{y}^*),\mathbf{y}\sim p(\mathbf{y})}[\max(0,\alpha+R_\phi(\mathbf{y})-R_\phi(\mathbf{y}^*))]$, where $p(\mathbf{y})$ is the action distribution and α is a hyperparameter.

Deep Q-Network, Inverse Reinforcement Learning and Human Preference Learning can all be represented in the form of Eq. (2). Once R_{ϕ^*} is obtained, it can be subsequently utilized to train an agent with standard RL techniques.

Imitation Learning Instead of learning a reward function, this line of methods trains the agent to mimic the aligned action, implicitly representing "*what we value*" [Torabi *et al.*, 2018]. Define a ground truth policy $\pi(\mathbf{y})$ and a learned policy $\pi_\theta(\mathbf{y})$ (agent) parameterized by θ, we could minimize the f-divergence[2] between the two policies [Go *et al.*, 2023]:

$$\theta^* = \underset{\theta}{\text{argmin}}\ D_f[\pi(\mathbf{y})||\pi_\theta(\mathbf{y})], \tag{3}$$

where $\pi(\mathbf{y})$ is the empirical distribution formed by a training set. Using KL-divergence, Eq. (3) becomes the traditional cross-entropy loss. This method directly learns an agent to produce behaviors aligned with humans' preferences/values. In Sec. 3, we will demonstrate how each popular paradigm of LLM alignment is connected with these two lines.

2.3 Big Model Alignment Goal and Evaluation

Alignment Goal. Before delving into *how to align*, we first briefly introduce *what to align with*. Discussions of alignment goal originate from the *Specification Problem*, *i.e.*, *how do we define the purpose we desire from AI?* [Leike *et al.*, 2018], which can be considered from two aspects [Gabriel, 2020]: (1) *normative aspect*: what goals we should encode into AI, and (2) the *technical one*: how do we formalize and model the goals. Failing to implement the goal might cause AI to seek loopholes and accomplish the objective in unintended ways, known as *Specification Gaming* [Skalse *et al.*, 2022]. From the former aspect, alignment goals range from instructions, intentions, preferences, to interest, values and so on [Gabriel, 2020]. Another popular goal is the *Helpful, Honest, and Harmless* (HHH) principle [Askell *et al.*, 2021]. However, a majority of work [Ouyang *et al.*, 2022; Rafailov *et al.*, 2023] emphasizes alignment approaches while ignoring analysis about what goal is the most appropriate. Misalignment inadvertently leads to unintended or undesirable harms and consequences [Casper *et al.*, 2023].

Alignment Evaluation. The evaluation of alignment refers to assessing how well an AI behaves in accordance with human intentions, namely calculating $\mathcal{L}$ in Eq (1). Early benchmarks assess AI's performance on specific risk criteria, such as toxicity, bias, and misinformation [Gehman *et al.*, 2020; Lin *et al.*, 2022]. Bai *et al.* [2022a] introduce a dataset comprising human preference data that assesses the helpfulness and harmlessness of AI. This could be measured by either the similarity (BLEU or ROUGE) between the generated context $\mathbf{y}$ and the good/bad reference $\mathbf{y}_w$/ $\mathbf{y}_l$, or by the reward given by a trained reward model $R_\theta(\mathbf{y})$ [Song *et al.*, 2023]. While similarity-based measurement is commonly used, it requires ground-truth references and results in low correlation with human judgments. Therefore, human evaluation is also involved [Wang *et al.*, 2022], despite being more time-consuming and costly. Recent studies endeavor to involve LLMs as AI evaluators in the process [Wang *et al.*, 2023], which is more efficient but suffers from inherent bias. There is potential for devising a framework that combines the advantages of automated and human evaluations[3].

2.4 The Challenges of Alignment

To achieve the alignment as defined in Sec. 1, various *Research Challenges* (RC) still need to be addressed. These challenges include, but are not limited to: RC1: *Alignment efficacy*. The performance of existing alignment methods requires improvement. How to align AI more accurately with desired goals without introducing unintended biases remains an open question. RC2: *Alignment Generalization*. Alignment goals might vary with time, culture, and context. It's essential to enable learned AI to keep aligned when deployed into diverse scenarios [de Font-Reaulx, 2022]. RC3: *Data and training efficiency*. Alignment training typically requires a substantial amount of manually annotated data, which is time- or labor-consuming and unable to keep pace with the rapid evolution of AI [Casper *et al.*, 2023]. RC4: *Interpretability of alignment*. Understanding and interpreting the alignment process and value-based decision-making of AI is essential for AI trust and further improvement, which is regarded as one of the 'biggest open questions' [Ouyang *et al.*, 2022]. RC5: *Alignment taxes*. Alignment could potentially hinder the capabilities of AI compared to its original counterpart [Askell *et al.*, 2021]. Minimizing such influence or finding a better trade-off is an inevitable issue. RC6: *Scalable oversight* [Bowman *et al.*, 2022]. It's challenging to effectively regulate and control AI models as they become much more powerful (superintelligence) than humans to prevent undesirable issues. RC7: *Specification Gaming*. Alignment goals are usually specified as an approximated proxy objective, much simpler than the real one, leading to unintended and potentially harmful side effects [Skalse *et al.*, 2022]. Besides, developing effective evaluation methods is also critical for alignment. These challenges remain unsolved and require more in-depth exploration from the community.

3 Alignment Methods

The alignment approaches for LLMs mainly fall into three paradigms (Fig. 2): RL-based Alignment (Sec. 3.1), SFT-based Alignment (Sec. 3.2), and Inference-Time Alignment (Sec. 3.3). In this section, we will introduce and discuss each of these approaches, as well as the LMM alignment, and establish their connections to the definition introduced in Sec. 2.

3.1 RL-based Alignment

The past two years have witnessed a prevalent alignment paradigm, Reinforcement Learning from Human Feedback (RLHF) [Ouyang *et al.*, 2022], which primarily belongs to *Value Learning* but can be also regarded as a combination of both lines in Sec. 2. Given a dataset D comprising prompts (instructions) $\mathbf{x}$ and manually labeled pairs of preferred and

[2] https://en.wikipedia.org/wiki/F-divergence

[3] More discussions on alignment goal: arxiv.org/abs/2308.12014

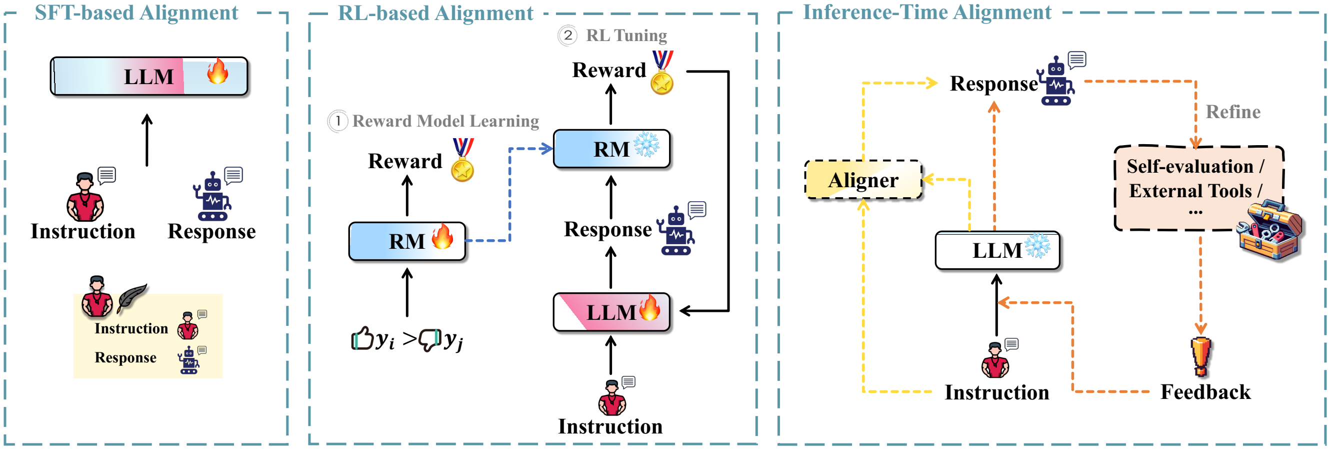


Figure 2: Illustrations of different alignment paradigms.

dispreferred model responses, $\mathbf{y}_w$ and $\mathbf{y}_l$, respectively, a typical RL alignment process consists of three steps:

(1) *Supervised fine-tuning* (SFT) step: Using

$$\mathcal{L}_{\text{SFT}}(\theta) = -\log \frac{1}{N}\sum_i \pi_\theta^{\text{SFT}}(\mathbf{y}^i|\mathbf{x}^i), \tag{4}$$

where N is the size of training data, to fine-tune the LLM π_θ to endow it with instruction-following capabilities. $\mathbf{y}^i$ is the collected human-written high-quality model response to $\mathbf{x}^i$, usually denoted as $\mathbf{y}_w$.

(2) *Reward Model Learning*: Training a reward model (RM) $R_\phi(r|\mathbf{y})$ from the preference data D which outputs a scalar reward r representing preferences learned from humans, by minimizing the following loss:

$$\mathcal{L}_{\text{RM}}(\phi) = -\mathbb{E}_D \log\left(\sigma\left(R_\phi\left(\mathbf{y}_w^i|\mathbf{x}_i\right) - R_\phi\left(\mathbf{y}_l^i|\mathbf{x}_i\right)\right)\right). \tag{5}$$

(3) *RL Tuning*: employing a policy-based deep RL algorithm, typically Proximal Policy Optimization (PPO), to optimize the LLM π_θ using the learned reward model with

$$\max_{\pi_\theta} \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D},\mathbf{y}\sim\pi_\theta} [R_\phi(\mathbf{y}|\mathbf{x})] - \lambda \text{KL}\left[\pi_\theta(\mathbf{y}|\mathbf{x})||\pi_{\text{SFT}}(\mathbf{y}|\mathbf{x})\right], \tag{6}$$

where λ is a parameter constraining the deviation from the original model π_{SFT}, *a.k.a. reference model*. This step maximizes the rewards obtained by the LLM, providing a feasible approach for learning human interaction and feedback.

Obviously, Eq. (5) is a kind of value learning that maximizes the margin between the ground truth action and the worse one. Taking a further step, we omit $\mathbf{x}$ and replace the sigmoid loss in Eq. (5) with a margin loss. By setting the ground truth reward r^* for all $\mathbf{y}_w$ and $\mathbf{y}_l$ to 1 and 0, respectively, we transform Eq. (5) into $\mathbb{E}_{p(\mathbf{y},r^*)}|r^* - R_\phi(\mathbf{y})|$, the form of Eq. (2). Furthermore, we could represent the reward as a delta distribution, and then the action-based reward learning can also be reformed as reward-based value learning, $\operatorname{argmin}_\phi \mathbb{E}_{p(\mathbf{y},r^*)}|r^* - R_\phi(\mathbf{y})| = \operatorname{argmin}_\phi \text{TV}[r^*(\mathbf{y})||R_\phi(\mathbf{y})]$, where TV is the Total Variation Distance.

In this way, by modifying the reward $R_\phi(\mathbf{y})$ in Eq. (6) as $\log R_\phi(\mathbf{y})$ and incorporating an entropy regularization for π_θ, we could unify the Reward Model Learning step and the RL Tuning one as f-divergence optimization:

$$\begin{aligned} \operatorname*{argmin}_{\phi,\theta} \underbrace{\text{TV}\left[r^*(\mathbf{y})||R_\phi(\mathbf{y})\right]}_{\text{Value Learning}} + \underbrace{\text{KL}\left[\pi_\theta(\mathbf{y})||R_\phi(\mathbf{y})\right]}_{\text{RL Tuning}} \\ + \underbrace{\lambda\text{KL}\left[\pi_\theta(\mathbf{y})||\pi_{\text{SFT}}(\mathbf{y})\right]}_{\text{Imitation Learning}}, \end{aligned} \tag{7}$$

where the first term models the reward, the second matches the policy with rewards, and the last one enforces the LLM to mimic its previous version to mitigate catastrophic forgetting.

The idea of RLHF was initially revealed in [Christiano *et al.*, 2017], where human preference was expressed over segments of agent trajectory for deep reinforcement learning, enabling the learning of more complex behaviors. After that, Stiennon *et al.* [2020] adapt the RLHF technique to the summarization task, learning human preferences on different summaries and resulting in a significant quality improvement. In addition, Nakano *et al.* [2021] propose WebGPT, which fine-tunes GPT-3 and employs RLHF to enhance web navigation and information retrieval capabilities. Such early studies using RLHF primarily aim to enhance model performance, specifically in terms of 'helpfulness' or 'honesty', potentially neglecting 'harmlessness' (HHH) [Askell *et al.*, 2021]. This omission might cause the misalignment between LLMs and human values, resulting in model outputs that are harmful or untruthful to users, as mentioned in Sec. 1. To reduce such harmful information, InstructGPT [Ouyang *et al.*, 2022] utilizes RLHF to align with the user's intentions, represented by the labeled model responses, to adhere to the HHH principle. RLHF technology directly gave rise to one of the most successful interactive dialogue LLMs, ChatGPT, igniting a spark toward Artificial General Intelligence (AGI).

Regardless of its satisfactory effectiveness, RLHF requires simultaneously loading at least three LLMs, namely π_θ, π_{SFT}, and R_ϕ, as well as a large amount of high-quality manually labeled data, $D(\mathbf{x}, \mathbf{y}_w, \mathbf{y}_l)$. This poses an unaffordable data/training cost (RC3). To tackle this challenge, Constitutional AI [Bai *et al.*, 2022b] was proposed to achieve alignment without human labels. This method is similar to RLHF but automatically creates the pairs $(\mathbf{y}_w, \mathbf{y}_l)$ by asking the LLM to generate and revise its responses. This framework facilitates a new line of alignment, namely *RL from AI Feedback* (RLAIF). Subsequently, different variants of RLAIF were developed. Kim *et al.* [2023] first train the reward model by utilizing synthetic preference data derived from LLMs with various scales and prompts. They then automatically generate high-quality demonstrations for the SFT step, followed by conducting RL tuning with the reward model. On the other hand, to improve the computational efficiency of RLHF, Gulcehre *et al.* [2023] propose an offline Reinforced Self-Training (ReST) method. ReST samples multiple responses from the latest LLM policy to augment the training dataset (Grow Step), and then uses the filtered data to fine-tune the LLM policy with an offline RL objective (Improve Step).

Pros and Cons: RLHF has proven to be effective in achieving relatively good generalization, holding the potential to better utilize human feedback signals. However, it is notorious for unstable training and high training/data cost (RC3), which impedes RLHF's further adaptability (RC2) and scalability (RC6). Besides, the trade-off between different terms in Eq. (7) is intractable (RC5), and RC4&7 also remain unresolved [Casper *et al.*, 2023].

3.2 SFT-based Alignment

To reduce the complexity and cost of alignment, researchers have paid more attention to the first step of RLHF, Supervised Fine-Tuning (SFT), and proposed a range of sophisticated SFT variations to reach the same performance as RLHF. Omitting $\mathbf{x}$ for brevity, a general form of SFT alignment is:

$$\begin{aligned}\underset{\theta}{\operatorname{argmin}}\ & -\mathbb{E}_{p(\mathbf{y}_w,\mathbf{y}_l)}\left[\log \pi_\theta(\mathbf{y}_w) - \log \pi_\theta(\mathbf{y}_l)\right] \\ & \propto \mathrm{KL}\left[p(\mathbf{y}_w)\|\pi_\theta(\mathbf{y}_w)\right] - \mathrm{KL}\left[p(\mathbf{y}_l)\|\pi_\theta(\mathbf{y}_l)\right], \end{aligned} \quad (8)$$

indicating that this paradigm is a member of imitation learning in Eq. (3), which directly learns to mimic the preferred behaviors while unlearning the dispreferred ones.

Without using negative examples $\mathbf{y}_l$, Eq. (8) reverts to conventional *instruction tuning*. For example, LIMA [Zhou *et al.*, 2023] assumes that an LLM's knowledge is primarily gained during pretraining, and alignment teaches the model which formats to use in interactions. It achieves the alignment of an LLaMA-65B model by utilizing a limited set of 1k meticulously curated instructions and their corresponding golden responses. Like RLAIF, such $(instruction, response)$ data could also be automatically constructed. Wang *et al.* [2022] propose SELF-INSTRUCT, a semi-automated method for generating instruction data to improve LLMs' instruction following capabilities. Similarly, SELF-ALIGN [Sun *et al.*, 2023], based on the SELF-INSTRUCT approach, incorporates additional human-defined value principles to generate more helpful, ethical, and reliable responses. To address the limitation of the methods above using only positive feedback $\mathbf{y}_w$, Chain of Hindsight (CoH) [Liu *et al.*, 2023a] was developed to utilize the paired feedback. During the training process, a prefix "Good" is appended to the preferred response, and "Bad" to $\mathbf{y}_l$. At inference, the LLM is instructed with "Good" to produce aligned responses. CoH is equivalent to learning a conditional policy $\pi_\theta(\mathbf{y}|r)$ conditioned on the reward r, and $r = 1$ ("Good") for $\mathbf{y}_w$ otherwise $r = 0$ ("Bad"), that is, $\mathbb{E}_{p(\mathbf{y},r)} \log \pi_\theta(\mathbf{y}|r) \propto \mathrm{KL}\left[p(\mathbf{y},r)\|\pi_\theta(\mathbf{y},r)\right]$. Even if aligned LLMs are trained to follow human values and avoid 'intentional' harms, they can still be susceptible to attacks from malicious users. To tackle this issue, Liu *et al.* [2022] propose SECOND THOUGHTS. This method first gets the unaligned source response and an aligned target response, and then makes the LLM learn to make edits to recover from a poisoned context during inference. As a result, even when provided with harmful context, the aligned LLM can generate content that aligns with human values, which is more robust to adversarial attacks. Besides directly learning ground truth actions, another line is to model the rank of responses, as ranking is often considered easier than scoring. Thus, the ranking-based loss is also incorporated into SFT alignment to capture the relative preferences and comparisons between different responses. Rank Responses to align Human Feedback (RRHF) [Yuan *et al.*, 2023] is one such method, which obtains a score for each response and then optimizes a ranking loss. This method makes the LLM learn to assign larger generation probabilities for responses with higher rewards.

From our analysis of RLHF in Sec. 3.1, we can see that value learning from target behaviors $\mathbf{y}$ can be transformed to the one from target rewards r. Analogously, imitation learning from behaviors can also be formed as reward learning. A milestone work in this line is Direct Preference Optimization (DPO) [Rafailov *et al.*, 2023]. This approach utilizes the Bradley-Terry (BT) preference model, $p^*(\mathbf{y}_w \succ \mathbf{y}_l|\mathbf{x}) = \frac{\exp(r^*(\mathbf{y}_w,\mathbf{x}))}{\exp(r^*(\mathbf{y}_w,\mathbf{x}))+\exp(r^*(\mathbf{y}_l,\mathbf{x}))}$, which models the probability that $\mathbf{y}_w$ is preferred than $\mathbf{y}_l$, to build a mapping between the optimal reward function and policy, $r^*(\mathbf{y},\mathbf{x}) \propto \lambda \log \frac{\pi^*(\mathbf{y}|\mathbf{x})}{\pi_{\mathrm{SFT}}(\mathbf{y}|\mathbf{x})}$, which is derived from the RLHF loss (Eq. (6)). This form allows the direct learning of the BT preference model by optimizing the LLM policy with the loss:

$$\mathcal{L}_{\mathrm{DPO}} = -\mathop{\mathbb{E}}_{\mathbf{x},\mathbf{y}_w,\mathbf{y}_l}\left[\log \sigma\left(\lambda \log \frac{\pi_\theta(\mathbf{y}_w|\mathbf{x})}{\pi_{\mathrm{SFT}}(\mathbf{y}_w|\mathbf{x})} - \lambda \log \frac{\pi_\theta(\mathbf{y}_l|\mathbf{x})}{\pi_{\mathrm{SFT}}(\mathbf{y}_l|\mathbf{x})}\right)\right]. \quad (9)$$

Note that DPO models human preference and implicitly represents the reward with policy, but we classify it into imitation learning, as the policy is still directly optimized using responses (actions). Following DPO, a series of preference modeling based SFT methods have emerged. Preference Ranking Optimization (PRO) [Song *et al.*, 2023] extends the BT preference model to capture the rank of multiple responses with the Plackett-Luce Model $p^*(\tau|\mathbf{y}_1,\dots,\mathbf{y}_K,\mathbf{x}) = \prod_{k=1}^{K} \frac{\exp(r^*(\mathbf{y}_{\tau(k)},\mathbf{x}))}{\sum_{j=k}^{K}\exp(r^*(\mathbf{y}_{\tau(j)},\mathbf{x}))}$, where K is the number of responses, τ is a permutation of these responses and $\tau(i)$ is the i-th response in the permutation. Furthermore, Azar *et al.* [2023] present ψPO objective for preference optimization, which unifies the RLHF and DPO methods. Moreover, they derived a specific variant of ψPO, the IPO method, to address the issue of overfitting by circumventing the BT preference model assumption with the training loss: $\mathcal{L}_{\mathrm{IPO}} = -\mathbb{E}_{(\mathbf{x},\mathbf{y}_w,\mathbf{y}_l)\sim D}[\log(\frac{\pi_\theta(\mathbf{y}_w|\mathbf{x})\pi_{\mathrm{SFT}}(\mathbf{y}_l|\mathbf{x})}{\pi_\theta(\mathbf{y}_l|\mathbf{x})\pi_{\mathrm{SFT}}(\mathbf{y}_w|\mathbf{x})}) - \frac{\lambda^{-1}}{2}]^2$.

Besides, inspired by contrastive learning, some methods learn patterns from positive samples that adhere to human expectations, while diverging from negative ones. Zhao *et al.* [2023b] apply the Sequence Likelihood Calibration (SLiC) method to effectively learn from human preferences (SLiC-HF). SLiC-HF includes a rank calibration loss and cross-entropy regularization term to encourage the model π_θ to generate positive sequences $\mathbf{y}_w$: $\mathcal{L}_{\text{SLiC}} = \max(0, \gamma - \log \pi_\theta(\mathbf{y}_w|\mathbf{x}) + \log \pi_\theta(\mathbf{y}_l|\mathbf{x})) - \lambda \log \pi_\theta(\mathbf{y}_{\text{ref}}|\mathbf{x})$, where $\mathbf{y}_{\text{ref}}$ is a regularization target, and γ and λ are hyper-parameters for margin and regularization weight, respectively. SLiC-HF uses a margin loss instead of the ratio loss in DPO. Liu *et al.* [2023c] first introduces a simulated human society called SANDBOX, which collects interaction data through communications among numerous LM-based social agents. Then based on contrastive learning, a novel alignment algorithm, Stable Alignment, is designed to learn *social alignment* from the collected data. Bhardwaj and Poria [2023] propose a RED-INSTRUCT method for achieving safety alignment in LLMs. The method involves constructing HARMFULQA using blue and red data. Then SAFE-ALIGN strategies are applied to fine-tune Vicuna, moving the model towards a safe and helpful response area in the distribution while steering it away from a harmful one. Hejna *et al.* [2023] propose Contrastive Preference Learning (CPL), which uses a regret-based model to learn a policy directly. By integrating the regret-based preference framework with the principle of Maximum Entropy (MaxEnt), the supervised objective of CPL can learn a consistent advantage function and converge to the optimal policy based on the expert's reward function.

Pros and Cons: SFT-based alignment provides a more flexible way to model human preference and improve alignment performance, corresponding to the imitation learning class introduced in Sec. 2. Compared to RLHF, SFT is much more efficient, requiring loading only one (Eq. (8)) or two (Eq. (9)) models. The training of SFT is more stable, and convergence is faster. However, since the value learning process is conducted in an implicit way, SFT alignment suffers from limited smoothness and generalization (RC2), and thus relatively poor performance (RC1). From Eq. (8), we can see that the imitation learning efficacy highly relies on the target behavior distribution being approximated, $p(\mathbf{y}_w), p(\mathbf{y}_l)$, imposing more stringent requirements on data quality (RC3). Besides, interpretability is worse, as the reward is not directly learned and hence hard to understand (RC4). Whether SFT can achieve or surpass the performance of RLHF one day is a question yet to be investigated.

3.3 Inference-Time Alignment

Considering the costs of SFT and RL, and the fact that most mainstream LLMs are black boxes, fine-tuning based alignment approaches become increasingly unaffordable or infeasible. Therefore, another popular paradigm, Inference-Time Alignment, which includes both an In-Context Learning (ICL) based method and a post-processing method, has attracted more attention. ICL leverages the massive knowledge and instruction-following capabilities of LLMs obtained during the pretraining and instruction tuning phases. By directly providing value instructions or K few-shot examples $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{K}$, ICL constrains the generation of the LLM to align with human values, avoiding the additional training. In fact, *ICL can also be regarded as a kind of imitation learning*. By incorporating a shared prompt concept [Xie *et al.*, 2021], $\mathbf{c}$, *e.g.*, values, minimizing the divergence between $p(\mathbf{y}, \mathbf{x}, \mathbf{c})$ and $\pi_\theta(\mathbf{y}, \mathbf{x}, \mathbf{c})$ can be transformed to optimizing:

$$\begin{aligned}
&\text{argmin KL}\left[p(\mathbf{y}, \mathbf{x}, \mathbf{c}) || \pi_\theta(\mathbf{y}, \mathbf{x}, \mathbf{c})\right] \\
=&\text{argmin } \mathbb{E}_{p(\mathbf{x},\mathbf{y})}\{\mathbb{E}_{p(\mathbf{c}|\mathbf{x},\mathbf{y})}\left[\log \pi_\theta(\mathbf{y}|\mathbf{x}, \mathbf{c})\right] \\
&\quad - \text{KL}\left[p(\mathbf{c}|\mathbf{x}, \mathbf{y}) || \pi_\theta(\mathbf{c}|\mathbf{x})\right]\}. \qquad (10)
\end{aligned}$$

Omitting the KL regularization term and freezing parameters θ, imitation learning can be viewed as implicit Bayesian inference, inferring the latent concept from given examples $\mathbf{x}, \mathbf{y}$, and driving the LLM to generate a connected response.

Concretely, the simplest way is to prompt LLMs to generate responses that adhere to human preferences [Ganguli *et al.*, 2023]. Han [2023] further retrieves relevant demonstration examples from SFT data, concatenating them with the input prompt. Lin *et al.* [2023] find that aligned LLMs primarily learn language styles matching human preferences, providing evidence in support of the "Superficial Alignment Hypothesis" [Zhou *et al.*, 2023]. Based on such findings, they propose to utilize three consistent stylistic examples and a system prompt for alignment. Considering the ever-changing and diverse human values in the real world, On-the-fly Preference Optimization (OPO) [Xu *et al.*, 2023] leverages Retrieval-Augmented Generation (RAG) to find context-aware values, achieving dynamical alignment. In addition, the post-processing method is also employed to achieve alignment. The generate-then-refine schema [Gou *et al.*, 2023] first generates initial responses and then enables LLMs to verify and rectify their own output. Rewindable Auto-regressive INference (RAIN) [Li *et al.*, 2023b] includes a self-evaluation mechanism to assess their own outputs and a rewind mechanism to search and rewind the token sets, serving as a plug-in module. Ji *et al.* [2024] streamline the alignment process through a copy and correction operation using Aligner. This approach necessitates only an additional module stacked onto the upstream LLM for alignment. The Aligner framework also facilitates weak-to-strong generalization.

Pros and Cons: Inference-Time Alignment (ITA) evades the need for training and labeled data (RC3). Without modifying the original model parameters, ITA avoids alignment tax (RC5) and proves more suitable for black-box models. Nonetheless, the efficacy of ITA heavily depends on the big model's ability to understand and follow instructions.

3.4 Multimodal Alignment

In addition to LLMs, Large Multimodal Models (LMMs) have also entered a new chapter of development in recent years, capable of processing multiple modalities simultaneously, such as images, videos, and texts, and learning mappings from one modality to another [Liu *et al.*, 2023b]. The initial achievements of aligning LLM indicate the potential for alignment in multimodal scenarios. In detail, a series of works integrate a pretrained vision encoder with an LLM and conduct instruction tuning to provide the LLM with visual QA capabilities, such as LLaVA [Liu *et al.*, 2023b], MiniGPT-4 [Zhu *et al.*, 2023], and so on [Li *et al.*, 2023a; Gong *et al.*, 2023; Dai *et al.*, 2023]. LLaVA [Liu *et al.*, 2023b] takes the first step in extending instruction tuning to LLMs, which combines the visual encoder of CLIP and an LLaMA-based language decoder, and conducts visual instruction tuning on a multimodal dataset generated by GPT-4. MiniGPT-4 [Zhu *et al.*, 2023] only trains a single projection layer to align the encoded visual features with the Vicuna language model. After instruction tuning on a curated small dataset, MiniGPT-4 can generate more natural and reliable language outputs. For text-to-image tasks, inspired by the effectiveness of RLHF in LLMs, Lee *et al.* [2023] propose a fine-tuning method for directly learning from human feedback. The process initially gathers human preference data about whether generated images correspond to their input text prompts, learns a reward model on this data, and finally, optimizes the text-to-image model using reward-weighted likelihood maximization to achieve alignment. To align with human aesthetic values, Wu *et al.* [2023] first utilize human-selected images to fine-tune the CLIP model as a preference classifier. Then this classifier is used to produce pseudo rewards for a training dataset, which is further employed to fine-tune the Stable Diffusion model. The trained model can generate images of better aesthetic quality that humans prefer.

Multimodality emerges as the future trajectory in big model advancements, providing a more direct avenue than language when engaging with humans. However, multimodal alignment is in its initial stages, focusing on aligning with human instructions but overlooking high-level and diverse human values like virtues and social norms. Ensuring harmlessness poses a significant and non-negligible challenge.

4 Further Challenges and Research

Ongoing and unexplored challenges. Most of the research challenges in Sec. 2 are still ongoing or totally unexplored, necessitating more detailed investigation. The community is currently mainly focusing on RC1 and RC3. Algorithm refinements, such as RLHF, DPO, and SLiC, are conducted to ensure that big models are aligned more accurately with desired behaviors and preferences. Studies of RLAIF focus on enhancing data efficiency by automating the generation of training data, thereby reducing human intervention and increasing scalability. Efforts are also made to improve training efficiency by simplifying RL-based methodologies, involving algorithms like DPO and RAIN, which quicken convergence and reduce GPU usage. Despite the progress and breakthrough so far, other problems such as the generalization (RC2, variability of values and context), interpretability (RC4, transparent alignment process and value-based reasoning), alignment tax (RC5, simultaneously minimizing LLM capability loss and maximizing alignment efficacy), scalable oversight (RC6, supervising and regulating superintelligence) and specification gaming (RC7, simple approximated proxy objective) represent critical future directions.

Measures to unresolved challenges. OpenAI has established a *Superalignment*[4] project, dedicating 20% of their computational resources to alignment challenges over the next four years. Their primary strategy, termed "*turning compute into alignment*," focuses on refining alignment iteratively through automated processes. The construction of an automated alignment researcher entails a tripartite process: 1) developing a scalable, AI-centric training methodology that guarantees both the generalization of the model (RC2) and the capacity for human oversight (RC6), 2) validating the system by devising methods for automatically detecting and interpreting problematic behaviors and internals to enhance robustness and interoperability (RC4), and 3) conducting stress tests to evaluate the pipeline's effectiveness by intentionally training misaligned models and verifying the detection of severe misalignments using their techniques. The final goal is to achieve the alignment of superintelligence [Nick, 2014].

Additionally, we should also *specify more appropriate alignment goals*. Fundamentally, big models aligned with human instructions or preferences still lack intrinsic knowledge of what constitutes truly "good" behaviors. This can lead to the specification gaming problem (RC7). It is essential to extend the objectives to aligning more coherently with human expectations, which might involve incorporating a deeper understanding of ethics, value theories from humanity and social science, and societal well-being into the alignment process. One promising direction for aligning big models is socialization alignment. The behaviors of social agents need to align with the specific values and norms of the society in which they interact with users in a iterative and dynamic manner. In this way, we can strive to create big models that not only perform actions preferred by humans but also align with broader notions of what is considered ethically good.

Further considerations of alignment. *Anthropic's core view*[5] categorizes alignment approaches into three scenarios according to the difficulty of improving AI safety. *Optimistic scenario*: the potential catastrophic risks from advanced AI due to safety failures are minimal, as existing techniques like RLHF [Ouyang *et al.*, 2022] and Constitutional AI [Bai *et al.*, 2022b] are deemed quite promising for alignment. The *Intermediate scenario* acknowledges the potential for catastrophic risks, necessitating substantial scientific and engineering efforts to counteract them, but remains achievable with dedicated endeavors. Lastly, the *Pessimistic scenario* posits AI safety as an unsolvable problem, arguing that controlling or dictating values to a system with greater intellectual capabilities than humans is impossible, thus opposes the development or deployment of highly advanced AI systems.

5 Conclusion

In this work, we delve into the origin and essence of alignment, systematically introducing its development, goals, formalization and evaluation. We also review existing work on alignment and analyze how each paradigm is derived from the original form and establish their intrinsic connections. By conducting a comprehensive analysis of alignment and identifying future challenges and research directions, we aim to contribute to the understanding and advancement of alignment approaches for big models, guiding these AI systems not only to avoid doing harm, but also to intent to do good, ultimately achieving a human-AI symbiotic future society.

[4] https://openai.com/blog/introducing-superalignment

[5] https://www.anthropic.com/index/core-views-on-ai-safety

Acknowledgements

*: X. Wang and S. Duan are co-first authors. †: X. Yi and J. Yao are corresponding authors. The work is partially supported by the National Nature Science Foundation of China (No. 62376199, 62076184, 62076182) and Shanghai Science and Technology Plan Project (No.21DZ1204800).

References

[Asilomar, 2017] A.I. Asilomar. Asilomar ai principles, 2017.

[Asimov, 1942] Isaac Asimov. Runaround, astouding science fiction. *New York*, 1942.

[Askell *et al.*, 2021] Amanda Askell, Yuntao Bai, Anna Chen, et al. A general language assistant as a laboratory for alignment. *ArXiv*, 2021.

[Aydin, 2023] Omer Aydin. Google bard generated literature review: Metaverse. *SSRN*, 2023.

[Azar *et al.*, 2023] Mohammad Gheshlaghi Azar, Mark Rowland, et al. A general theoretical paradigm to understand learning from human preferences. *ArXiv*, 2023.

[Bai *et al.*, 2022a] Yuntao Bai, Andy Jones, Kamal Ndousse, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *ArXiv*, 2022.

[Bai *et al.*, 2022b] Yuntao Bai, Saurav Kadavath, et al. Constitutional ai: Harmlessness from ai feedback. *ArXiv*, 2022.

[Bhardwaj and Poria, 2023] Rishabh Bhardwaj and Soujanya Poria. Red-teaming large language models using chain of utterances for safety-alignment. *ArXiv*, 2023.

[Bommasani *et al.*, 2021] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, et al. On the opportunities and risks of foundation models. *ArXiv*, 2021.

[Bostrom, 2003] Nick Bostrom. Ethical issues in advanced artificial intelligence. *Science fiction and philosophy: from time travel to superintelligence*, 2003.

[Bowman *et al.*, 2022] Samuel R Bowman, Jeeyoon Hyun, Ethan Perez, et al. Measuring progress on scalable oversight for large language models. *ArXiv*, 2022.

[Carroll, 2018] Micah Carroll. Overview of current ai alignment approaches. 2018.

[Casper *et al.*, 2023] Stephen Casper, Xander Davies, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *ArXiv*, 2023.

[Christiano *et al.*, 2017] Paul F. Christiano, Jan Leike, Tom B. Brown, et al. Deep reinforcement learning from human preferences. *NeurIPS*, 2017.

[Christiano, 2018] Paul Christiano. Clarifying ai alignment, 2018.

[Dai *et al.*, 2023] Wenliang Dai, Junnan Li, Dongxu Li, et al. Instructblip: Towards general-purpose vision-language models with instruction tuning. *ArXiv*, 2023.

[de Font-Reaulx, 2022] Paul de Font-Reaulx. Alignment as a dynamic process. *NeurIPS ML Safety Workshop*, 2022.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, et al. BERT: Pre-training of deep bidirectional transformers for language understanding. *AACL*, 2019.

[Everitt and Hutter, 2018] Tom Everitt and Marcus Hutter. The alignment problem for bayesian history-based reinforcement learners. *Under submission*, 2018.

[Gabriel, 2020] Iason Gabriel. Artificial intelligence, values, and alignment. *Minds and machines*, 2020.

[Ganguli *et al.*, 2023] Deep Ganguli, Amanda Askell, Nicholas Schiefer, et al. The capacity for moral self-correction in large language models. *ArXiv*, 2023.

[Gehman *et al.*, 2020] Samuel Gehman, Suchin Gururangan, et al. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. *EMNLP Findings*, 2020.

[Go *et al.*, 2023] Dongyoung Go, Tomasz Korbak, Germán Kruszewski, et al. Aligning language models with preferences through f-divergence minimization. *ArXiv*, 2023.

[Gong *et al.*, 2023] Tao Gong, Chengqi Lyu, Shilong Zhang, et al. Multimodal-gpt: A vision and language model for dialogue with humans. *ArXiv*, 2023.

[Gou *et al.*, 2023] Zhibin Gou, Zhihong Shao, Yeyun Gong, et al. Critic: Large language models can self-correct with tool-interactive critiquing. *ArXiv*, 2023.

[Gulcehre *et al.*, 2023] Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, et al. Reinforced self-training (rest) for language modeling. *ArXiv*, 2023.

[Hadfield-Menell *et al.*, 2016] Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, et al. Cooperative inverse reinforcement learning. *NeurIPS*, 2016.

[Han, 2023] Xiaochuang Han. In-context alignment: Chat with vanilla language models before fine-tuning. *ArXiv*, 2023.

[Hejna *et al.*, 2023] Joey Hejna, Rafael Rafailov, et al. Contrastive prefence learning: Learning from human feedback without rl. *ArXiv*, 2023.

[Horvitz and Selman, 2008] Eric Horvitz and Bart Selman. Aaai presidential panel on long-term ai futures: Interim report from the panel chairs. *AAAI*, 2008.

[Ji *et al.*, 2024] Jiaming Ji, Boyuan Chen, Hantao Lou, et al. Aligner: Achieving efficient alignment through weak-to-strong correction. *ArXiv*, 2024.

[Kaplan *et al.*, 2020] Jared Kaplan, Sam McCandlish, et al. Scaling laws for neural language models. *ArXiv*, 2020.

[Kenton *et al.*, 2021] Zachary Kenton, Tom Everitt, et al. Alignment of language agents. *ArXiv*, 2021.

[Kim *et al.*, 2023] Sungdong Kim, Sanghwan Bae, Jamin Shin, et al. Aligning large language models through synthetic feedback. *ArXiv*, 2023.

[Lee *et al.*, 2023] Kimin Lee, Hao Liu, et al. Aligning text-to-image models using human feedback. *ArXiv*, 2023.

[Leike *et al.*, 2018] Jan Leike, David Krueger, Tom Everitt, et al. Scalable agent alignment via reward modeling: a research direction. *ArXiv*, 2018.

[Li *et al.*, 2023a] Bo Li, Yuanhan Zhang, Liangyu Chen, et al. Otter: A multi-modal model with in-context instruction tuning. *ArXiv*, 2023.

[Li *et al.*, 2023b] Yuhui Li, Fangyun Wei, Jinjing Zhao, et al. Rain: Your language models can align themselves without finetuning. *ArXiv*, 2023.

[Lin *et al.*, 2022] Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. *ACL*, 2022.

[Lin *et al.*, 2023] Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, et al. The unlocking spell on base llms: Rethinking alignment via in-context learning. *ArXiv*, 2023.

[Liu *et al.*, 2022] Ruibo Liu, Chenyan Jia, Ge Zhang, et al. Second thoughts are best: Learning to re-align with human values from text edits. *NeurIPS*, 2022.

[Liu *et al.*, 2023a] Hao Liu, Carmelo Sferrazza, and Pieter Abbeel. Chain of hindsight aligns language models with feedback. *ArXiv*, 2023.

[Liu *et al.*, 2023b] Haotian Liu, Chunyuan Li, Qingyang Wu, et al. Visual instruction tuning. *ArXiv*, 2023.

[Liu *et al.*, 2023c] Ruibo Liu, Ruixin Yang, Chenyan Jia, et al. Training socially aligned language models in simulated human society. *ArXiv*, 2023.

[McKenzie *et al.*, 2023] Ian R McKenzie, Alexander Lyzhov, Michael Pieler, et al. Inverse scaling: When bigger isn't better. *ArXiv*, 2023.

[Mnih *et al.*, 2015] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, et al. Human-level control through deep reinforcement learning. *nature*, 2015.

[Nakano *et al.*, 2021] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, et al. Webgpt: Browser-assisted question-answering with human feedback. *ArXiv*, 2021.

[Nick, 2014] Bostrom Nick. Superintelligence: Paths, dangers, strategies. *Oxford University Press, Oxford*, 2014.

[Ouyang *et al.*, 2022] Long Ouyang, Jeffrey Wu, Xu Jiang, et al. Training language models to follow instructions with human feedback. *NeurIPS*, 2022.

[Rafailov *et al.*, 2023] Rafael Rafailov, Archit Sharma, Eric Mitchell, et al. Direct preference optimization: Your language model is secretly a reward model. *ArXiv*, 2023.

[Sheng *et al.*, 2019] Emily Sheng, Kai-Wei Chang, et al. The woman worked as a babysitter: On biases in language generation. *EMNLP*, 2019.

[Skalse *et al.*, 2022] Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, et al. Defining and characterizing reward gaming. *NeurIPS*, 2022.

[Soares and Fallenstein, 2014] Nate Soares and Benja Fallenstein. Aligning superintelligence with human interests: A technical research agenda. *MIRI technical report*, 2014.

[Song *et al.*, 2023] Feifan Song, Bowen Yu, et al. Preference ranking optimization for human alignment. *ArXiv*, 2023.

[Stiennon *et al.*, 2020] Nisan Stiennon, Long Ouyang, Jeffrey Wu, et al. Learning to summarize with human feedback. *NeurIPS*, 2020.

[Sun *et al.*, 2023] Zhiqing Sun, Yikang Shen, Qinhong Zhou, et al. Principle-driven self-alignment of language models from scratch with minimal human supervision. *ArXiv*, 2023.

[Team *et al.*, 2023] Gemini Team, Rohan Anil, Sebastian Borgeaud, et al. Gemini: a family of highly capable multimodal models. *ArXiv*, 2023.

[Torabi *et al.*, 2018] Faraz Torabi, Garrett Warnell, and Peter Stone. Behavioral cloning from observation. *ArXiv*, 2018.

[Touvron *et al.*, 2023] Hugo Touvron, Thibaut Lavril, Gautier Izacard, et al. Llama: Open and efficient foundation language models. *ArXiv*, 2023.

[Waldrop, 1987] M Mitchell Waldrop. A question of responsibility. *AI Magazine*, 1987.

[Wang *et al.*, 2022] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, et al. Self-instruct: Aligning language model with self generated instructions. *ArXiv*, 2022.

[Wang *et al.*, 2023] Yidong Wang, Zhuohao Yu, Zhengran Zeng, et al. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. *ArXiv*, 2023.

[Wei *et al.*, 2022] Jason Wei, Yi Tay, Rishi Bommasani, et al. Emergent abilities of large language models. *ArXiv*, 2022.

[Weidinger *et al.*, 2022] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, et al. Taxonomy of risks posed by language models. *ACM FAccT*, 2022.

[Wiener, 1960] Norbert Wiener. Some moral and technical consequences of automation: As machines learn they may develop unforeseen strategies at rates that baffle their programmers. *Science*, 1960.

[Wu *et al.*, 2023] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Better aligning text-to-image models with human preference. *ArXiv*, 2023.

[Xie *et al.*, 2021] Sang Michael Xie, Aditi Raghunathan, et al. An explanation of in-context learning as implicit bayesian inference. *ICLR*, 2021.

[Xu *et al.*, 2023] Chunpu Xu, Steffi Chern, Ethan Chern, et al. Align on the fly: Adapting chatbot behavior to established norms. *ArXiv*, 2023.

[Yuan *et al.*, 2023] Zheng Yuan, Hongyi Yuan, Chuanqi Tan, et al. Rrhf: Rank responses to align language models with human feedback without tears. *ArXiv*, 2023.

[Zhao *et al.*, 2023a] Wayne Xin Zhao, Kun Zhou, Junyi Li, et al. A survey of large language models. *ArXiv*, 2023.

[Zhao *et al.*, 2023b] Yao Zhao, Rishabh Joshi, Tianqi Liu, et al. Slic-hf: Sequence likelihood calibration with human feedback. *ArXiv*, 2023.

[Zhou *et al.*, 2023] Chunting Zhou, Pengfei Liu, Puxin Xu, et al. Lima: Less is more for alignment. *ArXiv*, 2023.

[Zhu *et al.*, 2023] Deyao Zhu, Jun Chen, et al. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *ArXiv*, 2023.