SOCIETAL ALIGNMENT FRAMEWORKS CAN IMPROVE LLM ALIGNMENT

Karolina Stańczak[1,2] **Nicholas Meade**[1,2] **Mehar Bhatia**[1,2] **Hattie Zhou**[1,3,4*]
Konstantin Böttinger[5] **Jeremy Barnes**[6] **Jason Stanley**[6] **Jessica Montgomery**[7]
Richard Zemel[8] **Nicolas Papernot**[9,10] **Nicolas Chapados**[1,6]
Denis Therien[2,6] **Timothy P Lillicrap**[10] **Ana Marasović**[11]
Sylvie Delacroix[12] **Gillian K Hadfield**[13] **Siva Reddy**[1,2,6]
[1]Mila - Quebec AI Institute [2]McGill University [3]Université de Montréal
[4]Anthropic [5]Fraunhofer AISEC [6]ServiceNow [7]University of Cambridge
[8]Columbia University [9]University of Toronto [10]Google DeepMind
[11]University of Utah [12]King's College London [13]Johns Hopkins University
karolina.stanczak@mila.quebec siva.reddy@mila.quebec

ABSTRACT

Recent progress in large language models (LLMs) has focused on producing responses that meet human expectations and align with shared values — a process coined *alignment*. However, aligning LLMs remains challenging due to the inherent disconnect between the complexity of human values and the narrow nature of the technological approaches designed to address them. Current alignment methods often lead to misspecified objectives, reflecting the broader issue of *incomplete contracts*, the impracticality of specifying a contract between a model developer, and the model that accounts for every scenario in LLM alignment. In this paper, we argue that improving LLM alignment requires incorporating insights from societal alignment frameworks, including social, economic, and contractual alignment, and discuss potential solutions drawn from these domains. Given the role of uncertainty in contract formalization within societal alignment frameworks, this paper investigates how it manifests in LLM alignment. We end our discussion by offering an alternative view on LLM alignment, framing the under-specified nature of its objectives as an opportunity rather than perfect their specification. Beyond technical improvements in LLM alignment, we discuss the need for participatory alignment interface designs.

1 INTRODUCTION

As large language models (LLMs) advance to unprecedented levels of proficiency in generating human-like language, aligning their behavior with human values has become a critical challenge to ensuring their usability in real-world applications (Leike et al., 2018; Gabriel, 2020; Ouyang et al., 2022; Shen et al., 2023). This alignment encompasses both *explicit* values, such as following instructions and being helpful, and *implicit* values, such as remaining truthful and avoiding biased or otherwise harmful outputs (Askell et al., 2021). In fact, the rise of LLM-based chat assistants has largely been driven by their ability to follow instructions and engage in open-ended dialogue, demonstrating the importance of alignment, enabled by algorithms such as reinforcement learning from human feedback (RLHF; Ouyang et al. 2022; Ziegler et al. 2020).

Despite these advancements, aligning LLMs with human values remains a formidable challenge (Wei et al., 2023; Williams et al., 2024; Greenblatt et al., 2024). This difficulty primarily stems from the fundamental gap between the intricacies of human values and the often narrow technological solutions (Hadfield-Menell & Hadfield, 2019). Current LLM alignment methods, such as RLHF, often result in misspecified alignment objectives, where reward functions reflect human values only within designer (or annotators) provided scenarios, a finite set among an infinite set of values, failing to generalize in unforeseen contexts (Amodei et al., 2016; Hadfield-Menell & Hadfield, 2019;

*Work done by HZ prior to joining Anthropic.

Skalse et al., 2024; Turner et al., 2020; 2021). While developers acknowledge the problem of misspecification (Leike et al., 2018; Shen et al., 2023), its root causes have been largely overlooked.

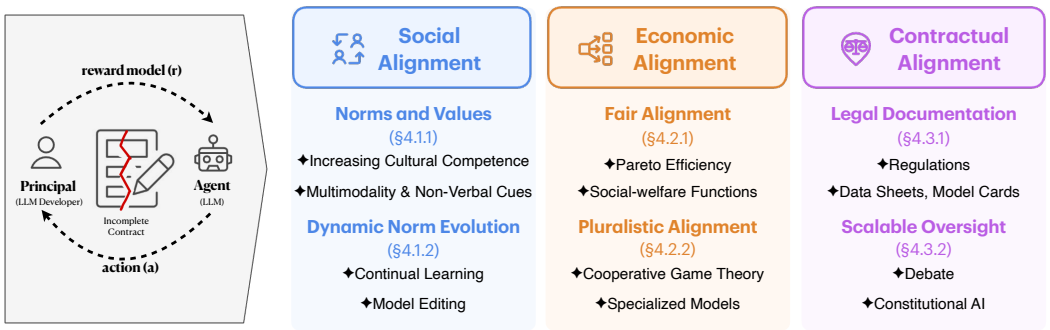

Figure 1: We view human-LLM interactions as a *principal-agent* framework, where a *principal* (a system designer) incentivizes an *agent* (an LLM) to take an action a by offering a reward r. This framework assumes that the agent's action is driven by its reward function, forming a pair (a, r) that serves as a *contract* between the agent and the principal. However, this contract is incomplete. To address this incompleteness, we explore societal alignment mechanisms of social, economic, and contractual alignment as guiding principles for LLM alignment.

To better understand this misalignment, we frame LLM alignment within a *principal-agent* framework (Eisenhardt, 1989), a well-established paradigm in economic theory. As shown in Figure 1, in this framework, LLM acts as the agent and the model developer (or user) serves as the principal. We define a *contract* as a pair: an action taken by the agent and the corresponding reward assigned by the principal. For example, a contract in LLM training could reward the model for generating responses that follow factual accuracy constraints while penalizing hallucinated outputs. The principal is able to steer the agent's behavior toward intended objectives with an appropriate reward. In an ideal scenario, a complete contract would perfectly align the agent's actions with the principal's objectives in all possible states of the world.

However, designing a fully specified contract that anticipates every possible scenario in model training is infeasible (Hadfield-Menell & Hadfield, 2019; Zhuang & Hadfield-Menell, 2020). In LLM alignment, this challenge is reflected in the reward function, which is derived from explicitly elicited values or implicitly implied values in the form of human preferences. Due to difficulties in quantifying complex human values (Leike et al., 2018), and the high annotation costs required to capture these values effectively (Klingefjord et al., 2024), the reward function is incomplete. As a result, it often leads to a solution that is detrimental to unspecified values (Zhuang & Hadfield-Menell, 2020).

These alignment challenges are not unique to LLMs. In fact, they echo broader alignment problems that humans encounter daily due to incomplete contracts. Institutions such as society, economy, and law enable us to thrive despite incompleteness. In this position piece, **we advocate for leveraging insights from societal alignment frameworks to guide the development of LLM alignment within incomplete contracting environments**. Drawing on principles from social alignment (Section 4.1), economic alignment (Section 4.2), and contractual alignment (Section 4.3), we propose solutions to guide behavior in incomplete contracting environments, much like they have for human societies (see Figure 1). However, even within these frameworks, uncertainty remains an inherent factor in incomplete contracting environments (Seita, 1984). In Section 5, we examine its real-world implications and how it manifests in LLM alignment. For instance, an LLM analyzing a patient's symptoms to suggest a diagnosis might lack access to the patient's full medical history or contextual background. In such cases, the model must navigate uncertainty to avoid overwhelming the user with complex, unfiltered medical information, which could lead to confusion or misinterpretation. Finally, we offer an alternative view on LLM alignment (Section 6), viewing the under-specified nature of its objectives as an opportunity rather than a flaw to be resolved solely through technological solutions. Instead, we discuss the need for participatory alignment interface designs that actively engage diverse stakeholders in LLM alignment.

2 CONTEMPORARY APPROACH TO LLM ALIGNMENT

Aligning LLMs with human values is commonly understood as training them to act in accordance with user intentions (Leike et al., 2018). The objective of LLM alignment is often conceptualized as fulfilling three core qualities, often referred to as the "3H" framework: honesty (regarding their capabilities, internal states, and knowledge), helpfulness (in performing requested tasks or answering questions within safe bounds), and harmlessness (encompassing both the refusal to fulfill harmful requests and the avoidance of generating harmful content) (Askell et al., 2021; Bai et al., 2022a).

A prominent approach to achieve this alignment is through a preference-based approach like RLHF. The RLHF pipeline usually includes three stages: supervised fine-tuning (SFT), preference sampling and reward model training (Christiano et al., 2017; Stiennon et al., 2020), and reinforcement learning fine-tuning either using proximal policy optimization (PPO; Schulman et al. 2017), or directly through policy optimization (DPO; Rafailov et al. 2023). The process usually starts with a generic pre-trained language model, which undergoes supervised learning on a high-quality dataset for specific downstream tasks. In this paper, we focus on the implications of the reward modeling stage due to its connection to an incomplete contract, which we will lay out in Section 3.

2.1 REWARD MODELING FROM HUMAN PREFERENCE.

In the reward modeling stage, for a given input prompt x, the SFT model generates paired outputs, $y_0, y_1 \in \mathcal{Y} \times \mathcal{Y}$, where $\mathcal{Y}$ denotes the set of all possible outputs that the model can generate in response to a given input. Human evaluators then select their preferred response, $y \in y_0, y_1$, providing data that guides the alignment process (Christiano et al., 2017; Stiennon et al., 2020). Human preferences are modeled probabilistically using frameworks like the Bradley-Terry model (Bradley & Terry, 1952). The preference probability for one response over another is expressed as:

$$p(y_1 \succ y_2 \mid x) = \frac{\exp(r(x, y_1))}{\exp(r(x, y_1)) + \exp(r(x, y_2))}, \tag{1}$$

where $r(x, y)$ is a latent reward function approximated by a parametric reward model, $r_\phi(x, y)$. Using a dataset of comparisons $\mathcal{D}$, the reward model is trained by minimizing the negative log-likelihood:

$$\mathcal{L}_R(r_\phi, \mathcal{D}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \Big[\log \sigma \big(r_\phi(x, y_w) - r_\phi(x, y_l) \big) \Big], \tag{2}$$

where σ is the logistic function and y_w and y_l denote the preferred and dispreferred completions among (y_1, y_2).

3 LLM ALIGNMENT AS A CONTRACT

In the following, we formalize LLM alignment through the lens of contract theory (Hadfield-Menell & Hadfield, 2019), a subfield of economics that studies how agreements are designed under conditions of incomplete information. We describe human-LLM interactions as a *principal-agent* relationship, where a *principal* (e.g., the user, system designer, or a company) seeks to incentivize an *agent* (an LLM) to act in a desired manner (Echenique et al., 2023) (see Figure 1). This framework provides a way to conceptualize how a model developer or the principal tries to align the agent's behavior with their objectives, using the agent's action and its reward function as a *contract*. In this section, we explore the contract formalization (Section 3.1) and how the incompleteness of this contract (Section 3.2) directly leads to misalignment (Section 3.3) in the context of LLM alignment.

3.1 CONTRACT FORMALIZATION

Following Echenique et al. (2023), we define a *contract* as a pair (a, r), where $a \in \mathcal{A}$ represents an action of an agent and $r : (\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}$ is a reward function.[1] The function r determines the agent's reward based on the observed input-output pair (x, y). In the context of a user-LLM interaction, an input $x \in \mathcal{X}$ corresponds to a user prompt, and output $y \in \mathcal{Y}$ is the LLM-generated response. A contract might be, for instance, a positive reward if the model avoids hate speech in

[1] Here we loosely refer a to mean one action or a series of actions that lead to an LLM output.

the output. Here, the reward function would be trained on prompt-response pairs, awarding higher scores to responses that do not contain hate speech.

The framework is initiated when the user, acting as the principal, initiates the interaction by prompting an LLM, thus implicitly proposing a contract. The LLM, acting as an agent, then implicitly either accepts or rejects this contract. Rejection of the contract displays in the LLM not converging towards the desired output, which is a generated response without hate speech. Upon implicitly accepting the contract, the LLM conducts an action a, which can be viewed as a probability distribution over all possible model outputs that satisfy the contract. We note that the user does not directly observe the LLM's internal decision of its action but only the resulting output y. Consequently, the agent is rewarded according to the agreed-upon reward function, $r(x, y)$, implemented as a reward signal during the training phase. The principal experiences the utility derived from the output y; that is, the user benefits from the generated response but also suffers if the model behaves adversarially.

3.2 The Challenge of Incomplete Contracting in AI

Although the specific implications of incomplete contracting for LLM alignment remain underexplored, the concept has been studied in the broader context of AI alignment (Hadfield-Menell & Hadfield, 2019). Alignment between the principal and the agent theoretically requires a *complete contract* (Williamson, 1973; Hadfield-Menell & Hadfield, 2019). A complete contract would perfectly align the principal's objectives with the agent's behavior in all possible states of the world. This requires that action a and reward function $r(x, y)$ be optimally defined for all input-output pairs. However, achieving complete contracts is practically infeasible for AI systems, rendering incomplete contracting unavoidable (Hadfield-Menell & Hadfield, 2019). This is primarily due to the fact that machine learning systems inherently operate with underspecified objectives (D'Amour et al., 2022), which stems from the practical difficulty in defining a reward function $r(x, y)$ that fully captures the complexities of the desired behavior.

The difficulty in specifying such a complete reward function arises from several issues. First, a key challenge for AI alignment generally, real-world applications are too complex to generate all possibilities, hindering the specification of every possible (a, r) pair (OpenAI, 2016). The space of possible outcomes, denoted by $\mathcal{Y}$ in the formalization is not tractable. This mirrors the challenge of LLM in generating outputs for new input it might receive during inference. The challenge extends beyond the practical limitations of fully specifying objectives. Second, even beyond these practical limitations, the challenge of translating complex human values into reward functions remains. Ambiguities in defining the desired action contribute to unintended and often undesirable outcomes.

3.3 Misalignment due to an Incomplete Contract

The incomplete contracting environment inherently leads to misalignment. This occurs when the reward function, $r(x, y)$ is underspecified and thus might incentivize outputs that diverge from the users's true objectives. A common outcome of reward misspecification is *reward hacking*, when an agent optimize for the reward itself instead of the intended behavior. For example, LLMs may exploit gaps in the specifications, such as in the "jailbreaking" phenomenon. Here, carefully crafted prompts elicit harmful responses from the model bypassing intended guardrails because their reward function is not specific enough, allowing the model to optimize without complying with safety requirements (Chao et al., 2024; Zou et al., 2023). Another example of reward hacking is an LLM trained to generate 'helpful' responses might learn to produce lengthy and verbose answers to prompts, as this might result in a higher score from the reward function even if it's not actually helpful to the user (Saito et al., 2023). A related issue occurs with "fake alignment," where the agents superficially comply with the training objective without necessarily having the intended internal goals (Greenblatt et al., 2024). Another challenge arises from the *inherent context dependence*. This involves designing reward functions that adapt appropriately to evolving contexts or requirements. A contract might specify desired behavior in a narrow scenario, but leave ambiguities for broader applications (Hadfield-Menell et al., 2017). For example, a contract that stipulates "no harmful bias" in a model is inherently underspecified since the definitions of "harmful" and "bias" are context-dependent.

4 SOCIETAL ALIGNMENT FRAMEWORKS

We present societal alignment frameworks, which can guide LLM alignment in the incomplete contracting environment. In the following, we discuss the alignment mechanisms of social theory (Section 4.1), economic theory (Section 4.2), and contractual theory (Section 4.3), and discuss potential solutions for improving the current LLM alignment.

4.1 SOCIAL ALIGNMENT

Human communication relies on a complex, largely implicit set of norms and values that help individuals interpret each other's intentions and the world around them (Bicchieri, 2017). However, this process is inherently ambiguous, as much of the meaning is conveyed implicitly rather than explicitly stated. Nonetheless, humans possess a unique ability called *normative competence*, which allows them to understand and judge whether certain behaviors are appropriate or inappropriate in a given context (Schutz, 1976). This shared understanding facilitates communication and fosters mutual understanding (Mercier & Sperber, 2017). A similar challenge arises in user-LLM interactions, where the absence of shared norms and values can result in misaligned outputs. Such information is often ingrained in cultures across the world (Hershcovich et al., 2022; Schäfer et al., 2015; Hanel et al., 2018). For example, an LLM providing evening activity recommendations without accounting for cultural context might suggest visiting a bar or consuming alcohol in a region where such activities are prohibited or socially unacceptable, leading to responses that fail to align with local norms and expectations. Incorporating societal norms and values into LLMs could equip them with mechanisms to interpret and adapt to human normative systems (Dragan et al., 2013), much like these aid alignment within human interactions (Bicchieri, 2017).

4.1.1 INSTILLING NORMS AND VALUES

While norms are specific rules that guide acceptable behavior in specific contexts, values are broader ideals representing overarching goals and aspirations, shaping what individuals strive for (Matsumoto, 2007). These can be instilled during LLM alignment in several ways.

As a fundamental tool of cooperative intelligence, language plays a crucial role in expressing and reinforcing both norms and values. LLMs, trained on vast datasets, absorb a multitude of signals about norms and values during training. However, while some attention has been given to broad ethical principles like helpfulness and harmlessness, an important aspect remains underexplored: "contextual rules" — human norms related to cultural conventions. These cntextual rules, while not directly influencing primary optimization objectives, are often followed due to tradition, or social norms. Despite their indirect nature, such rules can provide valuable signals about broader societal dynamics, thereby guiding the alignment of LLMs, as discussed in (Hadfield-Menell et al., 2019; Köster et al., 2020) within the broader context of AI alignment. Although efforts such as Ziems et al. (2022), Zhan et al. (2024), and Chiu et al. (2024) introduce datasets with collections of social norms, the influence of the collected norms on improving cultural adaptability in LLMs remains underexplored. Contextual rules could guide the style of language to align with cultural expectations. For instance, when interacting with users from diverse cultural backgrounds, LLM could account for cultural preferences by avoiding humor that might not translate well across cultures. However, existing models have been shown to predominantly reflect Western values, as they have been primarily trained on Western-centric data (Durmus et al., 2024), which limits their ability to represent multi-cultural values.

Human social norms and values are continuously shaped and evaluated through daily interactions with others. These interactions involve the exchange of multimodal signals, such as language, facial expressions, and gestures (Levinson & Holler, 2014). However, when interacting with LLMs, these cues are inherently absent, creating a normative gap in communication. Exploring multimodality for alignment — integrating non-verbal forms of communication such as visual, auditory, or behavioral signals — can serve as a promising line of research to address this normative void. By incorporating multimodal interactions, LLMs could better interpret and align with the implicit social expectations typically conveyed through non-verbal cues. This approach has the potential to enhance the contextual understanding and normative alignment of LLMs in diverse and complex human interactions.

4.1.2 ALLOWING FOR DYNAMIC NORMS AND VALUES

Norms and values are not static objects but dynamic equilibria that evolve through ongoing social interactions. The discursive practices that govern human interaction are normatively loaded and inherently dynamic. Norms are continuously re-articulated and negotiated within social contexts, evolving to address new challenges and cultural shifts (Gelfand et al., 2024). For instance, stereotypes are not fixed, as they emerge and transform over time. An example of an emerging stereotype is the shifting perception of remote work. Once seen as unprofessional or less productive, it is now widely accepted in many industries. If an LLM were trained primarily on pre-COVID data, it could reinforce outdated assumptions. While model editing and continual learning have been explored extensively for updating factual knowledge in LLMs (Mitchell et al., 2022; Benavides-Prado & Riddle, 2022), their application for adapting to evolving societal values and norms remains underexplored. Developing approaches to enable LLMs to identify, adapt to, and mitigate emerging biases dynamically is a crucial area for future research.

4.2 ECONOMIC ALIGNMENT

Economies rely on specialization and the division of labor, requiring coordination among groups of people to ensure efficient allocation of resources (Arrow, 1951). A central challenge in modern economic theory is aligning the interests of one actor with those of others (Hadfield-Menell & Hadfield, 2019). Similarly, aligning LLMs with diverse human values involves navigating trade-offs between individual and collective goals. Welfare economics complements this by focusing on developing optimization functions for resource allocation to maximize the objectives of an economic model. Additionally, a coherent social welfare objective function for LLMs cannot rely solely on subjective values. Instead, real-world implementations demand collective decisions about which values to prioritize (Arrow, 1951; Sen, 1985). Building from this, we explore strategies for integrating economic alignment frameworks to coordinate individual preferences, achieve collective objectives, and facilitate group-level aggregation, offering an alternative to imposing monolithic objective functions across diverse user groups.

4.2.1 ECONOMIC MECHANISMS FOR FAIR ALIGNMENT

LLMs must navigate group-level interactions, as societal challenges often center around collective behavior rather than individual actions. The alignment problem thus extends beyond aligning LLMs with individual values to addressing their role within groups. This includes prioritizing fairness by promoting values that ensure group stability and productivity while avoiding disruptions to collective decision-making processes.

In theoretical economics, perfect markets are often posited as achieving a Pareto-efficient distribution of welfare under a utilitarian framework (Arrow, 1951). Pareto efficiency, where no individual can be made better off without making someone worse off, is a benchmark for efficient resource allocation (Black et al., 2017). As shown in Boldi et al. (2024), Pareto efficiency offers a valuable lens for balancing competing human preferences and can serve as a foundation for techniques optimizing specific notions of group fairness, ensuring inclusive and equitable LLM alignment. Achieving such efficiency would mean tailoring the model's behavior to address diverse needs equitably, ensuring no group is disproportionately advantaged or disadvantaged without justification.

This problem has been investigated in the field of social welfare economics, where the aggregation of diverse preferences must be balanced to ensure the collective well-being of multiple groups (d'Aspremont & Gevers, 2002). For LLM alignment, these functions can guide the development of reward systems. As shown in general RLHF, developing welfare-centric objectives improves fairness (Pardeshi et al., 2024; Cousins et al., 2024).

4.2.2 ECONOMIC MECHANISMS FOR PLURALISTIC ALIGNMENT

Pluralistic alignment involves designing LLMs that can represent and respect a diverse set of human values and perspectives (Sorensen et al., 2024; Tanmay et al., 2023). Unlike monolithic approaches, which attempt to impose a singular objective function, pluralistic alignment embraces the complexity of modern societies. A critical aspect of LLM alignment involves determining how to elicit

and aggregate preferences when multiple humans are affected by the behavior of an artificial agent (Rossi et al., 2011; Rao et al., 2023; Conitzer et al., 2024).

The choice between developing general-purpose models and more specialized ones also shapes how uncertainty is handled. Specialized models tailored to specific contexts may better address immediate concerns but risk overlooking the broader cultural shifts that LLMs bring to ethically complex domains. For example, in healthcare or justice, specialized models can better align with local norms or regulatory frameworks. However, this approach risks fragmenting pluralistic values and creating inconsistencies across models. Cooperative game theory provides a framework to navigate these challenges by promoting fair resource allocation, fostering collaboration among stakeholders, and ensuring equitable outcomes (Chalkiadakis et al., 2011).

4.3 Contractual Alignment

Law-making and legal interpretation serve as mechanisms to translate opaque human goals and values into explicit, actionable directives. Legal scholars have long recognized the inherent impossibility of drafting complete contracts (Macneil, 1977; Williamson, 1973; Shavell, 1980; Maskin & Tirole, 1999; Tirole, 1999; Aghion & Holden, 2011). This limitation stems from several key challenges. First, certain states of the world are either unobservable or unverifiable, e.g., hiding assets in complex financial arrangements can be difficult for tax authorities to identify (Sears, 1921). Second, the bounded rationality of agents (i.e., humans) limits their ability to anticipate and optimize across the entire, combinatorially large space of potential scenarios (Williamson, 1973). Consequently, precisely computing optimal outcomes becomes intractable. Furthermore, the very description of all possible contingencies is often beyond human foresight, leading to loopholes in the design of rules (Katz, 2010). Even if feasible, the costs associated with drafting and enforcing fully specified contracts would likely be prohibitive. Given that these challenges are analogous to those encountered in aligning LLMs, where developers aim to ensure that models produce safe and correct outputs even for inputs not directly represented in training or alignment data, we investigate insights from contract theory as potential solutions to LLM alignment.

4.3.1 External Contractual Alignment

The formalization of contracts offers a framework for anticipating and specifying desired behaviors in human-LLM interactions (Jacovi et al., 2021). In this context, standardized documentation plays a crucial role in defining and communicating the LLMs' performance characteristics. Initiatives such as data statements, datasheets for datasets, model cards, reproducibility checklists, fairness checklists, and factsheets exemplify efforts to create clear, standardized guidelines that could inform the development of future regulations and legal frameworks for LLM alignment and data governance.

The rules that guide LLM alignment are currently largely constructed in consultation with domain and legal experts by adapting documents such as the UN Declaration of Human Rights (Anthropic, 2023a), through public input (Anthropic, 2023b), or in some cases, relying on designer instincts (Anthropic, 2023a; Solaiman & Dennison, 2021). Importantly, the European Commission has developed detailed guidelines for trustworthy AI, which provide a structured approach to ensuring that AI systems, including LLMs, adhere to ethical principles and societal norms.[2] These documents serve as critical tools for defining the terms of human-LLM contracts and offer a principled way to ensure that the view not only reflects the developer's personal views.

4.3.2 Internal Contractual Alignment

While the above discussion focused on aligning LLMs through external rules, another approach seeks to embed normative principles directly within the model's internal mechanisms. This approach is commonly known as 'constitutional AI,' wherein the LLM effectively develops an internal set of "principles" guiding its outputs. This involves transforming desired rules into the very training objectives of the LLM. These methods provide scalable oversight precisely because they move beyond the need for direct, case-by-case human intervention. Traditional preference-based training methods, such as collecting annotations on preferred and rejected outputs, aggregate

[2]The guidelines are available at `https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai/`.

multiple annotators' judgments into a shared standard, but they still require extensive human effort at scale. (Shen et al., 2023; Amodei et al., 2016). In contrast, scalable oversight approaches enable humans to oversee LLMs to manage complex tasks using structured mechanisms that generalize beyond individual preferences (Shen et al., 2023).

A promising scalable oversight method, debate (Irving et al., 2018; Irving & Askell, 2019), moves beyond aligning models to individual preferences by instilling structured reasoning. LLMs propose answers, engage in adversarial discussions, and refine arguments, with a human judge selecting the most well-supported response. On the other hand, in Constitutional AI, an LLM is guided by a concise constitution consisting of high-level principles (e.g., promoting fairness or avoiding harm) (Bai et al., 2022b; Sun et al., 2023). This constitution provides the basis for generating synthetic comparison examples, which are then used to fine-tune the LLM's policy. These methods were predominantly used to integrate human values but have the potential to enforce norms and regulations.

5 Societal Alignment Frameworks and their View on Uncertainty

By framing LLM alignment as a problem of contractual incompleteness and analyzing it through the lens of societal alignment frameworks, we observe that these frameworks recognize establishing contracts, much like alignment, as inherently uncertain (Seita, 1984). In the following, we examine uncertainty in the specific case of LLM alignment through the lens of societal alignment frameworks.

Prior research has identified epistemic uncertainty as one of the main challenges in LLM development (Shorinwa et al., 2024). This form of uncertainty arises from gaps in the model's knowledge, leading to uncertainty about factual information (Shorinwa et al., 2024; Jiang et al., 2021; Yadkori et al., 2024). Even aligned models remain susceptible to epistemic uncertainty, often failing to recognize their own knowledge limitations (Shorinwa et al., 2024). This inability to calibrate confidence scores to actual knowledge reflects a fundamental limitation of current LLM architectures. However, uncertainty in aligned LLMs presents additional complexity. The conversational nature of LLMs often creates an illusion of omniscience, making it difficult for users to discern the model's uncertainty (Delacroix, 2024). Furthermore, human interaction with models, combined with their in-context learning capabilities (Brown et al., 2020), allows users to provide task-specific context that can inadvertently bypass safety guardrails and mitigations implemented during training. Thus, models may leak unsafe information or performing harmful actions despite their intended safeguards (Glukhov et al., 2024).

While the unwanted epistemic uncertainty can undermine the reliability of LLMs, certain types of uncertainty are not only unavoidable but essential for their ethical deployment (Delacroix, 2024). This essential uncertainty can arise from evolving human values, conflicting societal norms, and the difficulty of translating abstract principles into model behavior. Aligning models to navigate trade-offs, such as between helpfulness and harmlessness or accuracy and fairness, requires addressing conflicting and often underspecified priorities, which introduces another source of uncertainty (Zollo et al., 2024; Yaghini et al., 2023). For instance, when deploying an LLM, we often want to maximize performance subject to some constraints or guardrails on behavior, e.g., chatbot should give users their desired output, as long as it is not too toxic. The effectiveness of balancing these conflicting priorities and the unintended consequences are often difficult to predict. However, this balancing act is also essential because it allows models to operate within complex, context-dependent environments where rigid adherence to a single objective could lead to harmful outcomes.

5.1 Uncertainty Communication

Building on the above, the inherent uncertainty in LLM alignment is not a weakness but often a valuable feature that enables models to handle complex scenarios ethically (Delacroix, 2024). In fact, as highlighted by Bhatt et al. (2021), uncertainty communication can be useful for obtaining fairer models by revealing data biases, improving decision-making by guiding reliance on predictions, and building trust in automated systems. Therefore, it is essential to develop methods for communicating uncertainty to users. Unlike humans, however, LLMs lack the non-verbal and contextual cues that naturally support communication (Bisconti, 2021). Existing research has shown that LLMs struggle to convey their uncertainty users, both implicitly (e.g., hedging language) and explicitly (e.g., con-

fidence scores), a skill that humans poses intuitively (Alkaissi & McFarlane, 2023; Liu et al., 2024; Shorinwa et al., 2024). On the other hand, humans themselves have varying levels of understanding regarding probability and statistics, which are needed to interpret model uncertainty estimates (Bhatt et al., 2021; Galesic & Garcia-Retamero, 2010). Furthermore, human cognition is subject to biases that can impede accurate interpretation of uncertainty (Kahneman, 2011; Reyna & Brainerd, 2008). These challenges can be partially addressed by choosing the appropriate communication methods, a key consideration for the design of effective user interfaces (Hullman et al., 2019), and by designing collaborative interaction environments, as discussed by Montemayor (2021).

6 Alternative View: The Democratic Opportunity Inherent in the Under-specified Nature of LLMs' Objectives

The challenge of aligning LLMs is often framed as a technical problem, one that can be solved through better reward modeling, training objectives, or oversight mechanisms. However, alignment is not merely a technological issue. It is fundamentally a societal one. Humans continuously navigate and redefine social norms, often through dialogue that refines our intuitions and moral expectations. These evolving dialogues shape our moral and social expectations, which, in turn, influence the values that guide our decision-making. The fact that these values change and often clash, is a good sign — a sign of ongoing critical engagement and willingness to question existing norms.

Since an LLM can be used as a conversational partner, the feedback given as a context can be leveraged to refine its behavior. Given the inherently dynamic nature of the values that inform education, healthcare, or justice practices, the key problem is to establish how to structure this feedback process. Different groups of users will evolve different values over time. Are there ways of incentivizing collective engagement with LLMs? Can bottom-up, iterative refinements be configured to support users in defining the values that preside over their practices (Delacroix, 2024)?

Framing LLM alignment as an incomplete contracting problem may be seen as an oversimplification of the complex socio-political dynamics. The contract metaphor, as discussed by Goldoni & Wilkinson (2018), reduces alignment to a straightforward agreement between stakeholders, neglecting the broader socio-political forces, conflicting norms, and inherent tensions that shape such systems. This technology-centric framing risks misrepresenting alignment's pluralistic nature. While societal alignment frameworks aim to address these issues, they too often rely on oversimplified assumptions. Beyond the issues with the contract metaphor, the focus on incompleteness (i.e., information asymmetries between the principal/agent) frames alignment as an epistemic designer-centric issue, rather than recognizing it first and foremost as a political question (Terzis, 2024). Given LLMs' unavoidable, normative effect on the practices within which they are deployed, the under-specified nature of LLMs' objectives presents an opportunity — not to perfect our specification methods, but to democratize the very process of determining what LLMs should optimize for.

The implications of this reframing extend to both research and practice. It suggests that alongside technical work such as reward modeling, we need equally sophisticated work on participatory interface designs. This dual focus acknowledges that effective participation requires not just theoretical frameworks for inclusion, but also concrete mechanisms through which diverse stakeholders can meaningfully shape LLM development (Kirk et al., 2024). This might include developing new methodologies for collective value articulation (Bergman et al., 2024), creating institutional structures for meaningful public participation in LLM development, and establishing mechanisms for ongoing societal oversight and input into LLMs' objectives and constraints.

7 Conclusions

In this paper, we argue that LLM alignment can be viewed through the lense of contract theory. Using a principal-agent model, where the principal (user or developer) defines a contract specifying the LLM's action and reward, we draw parallels between societal and LLM alignment challenges. While contract theory offers formalization tools, societal alignment instills norms, economic alignment addresses group coordination, and contractual alignment regulates behavior through legal and oversight mechanisms. Finally, we advocate for shifting from a developer-centered to a collaborative, user-centric, and iterative alignment approach.

ACKNOWLEDGEMENTS

This paper originated from the Bellairs Invitational Workshop on Contemporary, Foreseeable, and Catastrophic Risks of Large Language Models in April 2024. We thank all workshop participants for their valuable discussions and contributions.

REFERENCES

Philippe Aghion and Richard Holden. Incomplete contracts and the theory of the firm: What have we learned over the past 25 years? *Journal of Economic Perspectives*, 25(2):181–97, June 2011. doi: 10.1257/jep.25.2.181. URL https://www.aeaweb.org/articles?id=10.1257/jep.25.2.181.

Hussam Alkaissi and Samy I McFarlane. Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus*, 15, 2023. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC9939079/.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety, 2016. URL https://arxiv.org/abs/1606.06565.

Anthropic. Claude's constitution. https://www.anthropic.com/index/claudes-constitution, May 9 2023a. Accessed: 2025-01-27.

Anthropic. Collective constitutional AI: Aligning a language model with public input. https://www.anthropic.com/index/collective-constitutional-ai-aligning-a-language-model-with-public-input, October 17 2023b. Accessed: 2025-01-27.

Kenneth J Arrow. An extension of the basic theorems of classical welfare economics. In *Proceedings of the second Berkeley symposium on mathematical statistics and probability*, volume 2, pp. 507–533. University of California Press, 1951. URL https://projecteuclid.org/ebooks/berkeley-symposium-on-mathematical-statistics-and-probability/Proceedings-of-the-Second-Berkeley-Symposium-on-Mathematical-Statistics-and/chapter/An-Extension-of-the-Basic-Theorems-of-Classical-Welfare-Economics/bsmsp/1200500251.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. A general language assistant as a laboratory for alignment, 2021. URL https://arxiv.org/abs/2112.00861.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova Dassarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, John Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Christopher Olah, Benjamin Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback. *ArXiv*, abs/2204.05862, 2022a. URL https://arxiv.org/abs/2204.05862.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional AI: Harmlessness from AI feedback, 2022b. URL https://arxiv.org/abs/2212.08073.

Diana Benavides-Prado and Patricia Riddle. A theory for knowledge transfer in continual learning. In Sarath Chandar, Razvan Pascanu, and Doina Precup (eds.), *Proceedings of The 1st Conference on Lifelong Learning Agents*, volume 199 of *Proceedings of Machine Learning Research*, pp. 647–660. PMLR, 22–24 Aug 2022. URL https://proceedings.mlr.press/v199/prado22a.html.

Stevie Bergman, Nahema Marchal, John Mellor, Shakir Mohamed, Iason Gabriel, and William Isaac. STELA: A community-centred approach to norm elicitation for AI alignment. *Scientific Reports*, 14:6616, 2024. doi: 10.1038/s41598-024-56648-4. URL https://doi.org/10.1038/s41598-024-56648-4.

Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q. Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, Lama Nachman, Rumi Chunara, Madhulika Srikumar, Adrian Weller, and Alice Xiang. Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, pp. 401–413, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384735. doi: 10.1145/3461702.3462571. URL https://doi.org/10.1145/3461702.3462571.

Cristina Bicchieri. *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms*. Oxford University Press, 02 2017. ISBN 9780190622046. doi: 10.1093/acprof:oso/9780190622046.001.0001. URL https://doi.org/10.1093/acprof:oso/9780190622046.001.0001.

Paola Bisconti. How robots' unintentional metacommunication affects human–robot interactions: A systemic approach. *Minds & Machines*, 31:487–504, 2021. doi: 10.1007/s11023-021-09584-5. URL https://doi.org/10.1007/s11023-021-09584-5.

John D. Black, Nigar Hashimzade, and Gareth Myles (eds.). *A Dictionary of Economics*. Oxford University Press, Oxford, 5th edition, 2017. URL https://books.google.ca/books?id=WyvYDQAAQBAJ&pg=PT459&redir_esc=y#v=onepage&q&f=false.

Ryan Boldi, Li Ding, Lee Spector, and Scott Niekum. Pareto-optimal learning from preferences with hidden context, 2024. URL https://arxiv.org/abs/2406.15599.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. ISSN 00063444, 14643510. URL http://www.jstor.org/stable/2334029.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Georgios Chalkiadakis, Edith Elkind, and Michael Wooldridge. *Computational Aspects of Cooperative Game Theory (Synthesis Lectures on Artificial Inetlligence and Machine Learning)*. Morgan & Claypool Publishers, 1st edition, 2011. ISBN 1608456528. URL https://link.springer.com/book/10.1007/978-3-031-01558-8.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries, 2024. URL https://arxiv.org/abs/2310.08419.

Yu Ying Chiu, Liwei Jiang, Maria Antoniak, Chan Young Park, Shuyue Stella Li, Mehar Bhatia, Sahithya Ravi, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. CulturalTeaming: AI-assisted interactive red-teaming for challenging LLMs'(lack of) multicultural knowledge. *arXiv preprint arXiv:2404.06664*, 2024.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf.

Vincent Conitzer, Rachel Freedman, Jobst Heitzig, Wesley H. Holliday, Bob M. Jacobs, Nathan Lambert, Milan Mosse, Eric Pacuit, Stuart Russell, Hailey Schoelkopf, Emanuel Tewolde, and William S. Zwicker. Position: Social choice should guide AI alignment in dealing with diverse human feedback. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 9346–9360. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/conitzer24a.html.

Cyrus Cousins, Kavosh Asadi, Elita Lobo, and Michael Littman. On welfare-centric fair reinforcement learning. *Reinforcement Learning Journal*, 3:1124–1137, 2024. URL https://rlj.cs.umass.edu/2024/papers/Paper133.html.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(1), January 2022. ISSN 1532-4435. URL https://dl.acm.org/doi/pdf/10.5555/3586589.3586815.

Claude d'Aspremont and Louis Gevers. Social welfare functionals and interpersonal comparability. In *Handbook of Social Choice and Welfare*, volume 1 of *Handbook of Social Choice and Welfare*, pp. 459–541. Elsevier, 2002. doi: https://doi.org/10.1016/S1574-0110(02)80014-5. URL https://www.sciencedirect.com/science/article/pii/S1574011002800145.

Sylvie Delacroix. Lost in conversation? Hermeneutics, uncertainty and large language models. *SSRN*, April 21 2024. URL https://ssrn.com/abstract=4751774.

Anca D. Dragan, Kenton C.T. Lee, and Siddhartha S. Srinivasa. Legibility and predictability of robot motion. In *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction*, HRI '13, pp. 301–308. IEEE Press, 2013. ISBN 9781467330558. URL https://ieeexplore.ieee.org/document/6483603.

Esin Durmus, Karina Nguyen, Thomas Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. Towards measuring the representation of subjective global opinions in language models. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=zl16jLb91v.

F. Echenique, N. Immorlica, V.V. Vazirani, and A.E. Roth. *Contract Theory*, pp. 614–624. Cambridge University Press, 2023. ISBN 9781108831994. URL https://books.google.ca/books?id=1ea-EAAAQBAJ.

Kathleen M. Eisenhardt. Agency theory: An assessment and review. *The Academy of Management Review*, 14(1):57–74, 1989. ISSN 03637425. URL http://www.jstor.org/stable/258191.

Iason Gabriel. Artificial intelligence, values, and alignment. *Minds & Machines*, 30(3):411–437, 2020. doi: 10.1007/s11023-020-09539-2. URL https://doi.org/10.1007/s11023-020-09539-2.

Mirta Galesic and Rocio Garcia-Retamero. Statistical numeracy for health: A cross-cultural comparison with probabilistic national samples. *Archives of Internal Medicine*, 170(5):462–468, 03 2010. ISSN 0003-9926. doi: 10.1001/archinternmed.2009.481. URL https://doi.org/10.1001/archinternmed.2009.481.

Michele J. Gelfand, Sergey Gavrilets, and Nathan Nunn. Norm dynamics: Interdisciplinary perspectives on social norm emergence, persistence, and change. *Annual Review of Psychology*, 75 (Volume 75, 2024):341–378, 2024. ISSN 1545-2085. doi: https://doi.org/10.1146/annurev-psych-033020-013319. URL https://www.annualreviews.org/content/journals/10.1146/annurev-psych-033020-013319.

David Glukhov, Ziwen Han, Ilia Shumailov, Vardan Papyan, and Nicolas Papernot. Breach by a thousand leaks: Unsafe information leakage in 'safe' AI responses, 2024. URL https://arxiv.org/abs/2407.02551.

Marco Goldoni and Michael A. Wilkinson. The material constitution. *The Modern Law Review*, 81 (4):567–597, 2018. ISSN 00267961, 14682230. URL http://www.jstor.org/stable/26647134.

Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, Akbir Khan, Julian Michael, Sören Mindermann, Ethan Perez, Linda Petrini, Jonathan Uesato, Jared Kaplan, Buck Shlegeris, Samuel R. Bowman, and Evan Hubinger. Alignment faking in large language models, 2024. URL https://arxiv.org/abs/2412.14093.

Dylan Hadfield-Menell and Gillian K. Hadfield. Incomplete contracting and AI alignment. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, pp. 417–422, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450363242. doi: 10.1145/3306618.3314250. URL https://doi.org/10.1145/3306618.3314250.

Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, Stuart J Russell, and Anca Dragan. Inverse reward design. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/32fdab6559cdfa4f167f8c31b9199643-Paper.pdf.

Dylan Hadfield-Menell, Mckane Andrus, and Gillian Hadfield. Legible normativity for AI alignment: The value of silly rules. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, pp. 115–121, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450363242. doi: 10.1145/3306618.3314258. URL https://doi.org/10.1145/3306618.3314258.

Paul H. P. Hanel, Gregory R. Maio, Ana K. S. Soares, Katia C. Vione, Gabriel L. de Holanda Coelho, Valdiney V. Gouveia, Appasaheb C. Patil, Shanmukh V. Kamble, and Antony S. R. Manstead. Cross-cultural differences and similarities in human value instantiation. *Frontiers in Psychology*, 9, 2018. ISSN 1664-1078. doi: 10.3389/fpsyg.2018.00849. URL https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2018.00849.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. Challenges and strategies in cross-cultural NLP. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6997–7013, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.482. URL https://aclanthology.org/2022.acl-long.482/.

Jessica Hullman, Xiaoli Qiao, Michael Correll, Alex Kale, and Matthew Kay. In pursuit of error: A survey of uncertainty visualization evaluation. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):903–913, 2019. doi: 10.1109/TVCG.2018.2864889. URL https://ieeexplore.ieee.org/document/8457476.

Geoffrey Irving and Amanda Askell. AI safety needs social scientists. *Distill*, 2019. doi: 10.23915/distill.00021. URL https://distill.pub/2019/safety-needs-social-scientists/.

Geoffrey Irving, Paul Christiano, and Dario Amodei. AI safety via debate, 2018. URL https://arxiv.org/abs/1805.00899.

Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pp. 624–635, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445923. URL https://doi.org/10.1145/3442188.3445923.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. How can we know when language models know? On the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977, 2021. doi: 10.1162/tacl_a_00407. URL https://aclanthology.org/2021.tacl-1.57/.

Daniel Kahneman. *Thinking, Fast and Slow*. Allen Lane, 2011. URL https://books.google.ca/books?id=AV9x8XakdV0C.

Leo Katz. A theory of loopholes. *The Journal of Legal Studies*, 39(1):1–31, 2010. ISSN 00472530, 15375366. URL http://www.jstor.org/stable/10.1086/649046.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Michael Bean, Katerina Margatina, Rafael Mosquera, Juan Manuel Ciro, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. The PRISM alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=DFr5hteojx.

Oliver Klingefjord, Ryan Lowe, and Joe Edelman. What are human values, and how do we align AI to them?, 2024. URL https://arxiv.org/abs/2404.10636.

Raphael Köster, Dylan Hadfield-Menell, Gillian K. Hadfield, and Joel Z. Leibo. Silly rules improve the capacity of agents to learn stable enforcement and compliance behaviors. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '20, pp. 1887–1888, Richland, SC, 2020. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450375184. URL https://dl.acm.org/doi/10.5555/3398761.3399016.

Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling: A research direction, 2018. URL https://arxiv.org/abs/1811.07871.

Stephen C. Levinson and Judith Holler. The origin of human multi-modal communication. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1651):20130302, 2014. doi: 10.1098/rstb.2013.0302. URL https://royalsocietypublishing.org/doi/abs/10.1098/rstb.2013.0302.

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, 2024. URL https://arxiv.org/abs/2308.05374.

Ian R Macneil. Contracts: Adjustment of long-term economic relations under classical, neoclassical, and relational contract law. *Northwestern University Law Review*, 72:854, 1977. URL https://heinonline.org/HOL/LandingPage?handle=hein.journals/illlr72&div=46&id=&page=.

Eric Maskin and Jean Tirole. Unforeseen Contingencies and Incomplete Contracts. *The Review of Economic Studies*, 66(1):83–114, 01 1999. ISSN 0034-6527. doi: 10.1111/1467-937X.00079. URL https://doi.org/10.1111/1467-937X.00079.

David Matsumoto. Culture, context, and behavior. *Journal of Personality*, 75(6):1285–1320, 2007. doi: https://doi.org/10.1111/j.1467-6494.2007.00476.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-6494.2007.00476.x.

Hugo Mercier and Dan Sperber. *The Enigma of Reason*. Harvard University Press, 2017. ISBN 9780674368309. URL http://www.jstor.org/stable/j.ctv2sp3dd8.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D Manning, and Chelsea Finn. Memory-based model editing at scale. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 15817–15831. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/mitchell22a.html.

Carlos Montemayor. Language and intelligence. *Minds & Machines*, 31:471–486, 2021. doi: 10.1007/s11023-021-09568-5. URL https://doi.org/10.1007/s11023-021-09568-5.

OpenAI. Faulty reward functions in the wild. https://openai.com/index/faulty-reward-functions/, 2016. Accessed: 2025-01-10.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

Kanad Shrikar Pardeshi, Itai Shapira, Ariel D. Procaccia, and Aarti Singh. Learning social welfare functions. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=7O6KtaAr8n.

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. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=HPuSIXJaa9.

Abhinav Sukumar Rao, Aditi Khandelwal, Kumar Tanmay, Utkarsh Agarwal, and Monojit Choudhury. Ethical reasoning over moral alignment: A case and framework for in-context ethical policies in LLMs. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 13370–13388, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.892. URL https://aclanthology.org/2023.findings-emnlp.892/.

Valerie F. Reyna and Charles J. Brainerd. Numeracy, ratio bias, and denominator neglect in judgments of risk and probability. *Learning and Individual Differences*, 18(1):89–107, 2008. ISSN 1041-6080. doi: https://doi.org/10.1016/j.lindif.2007.03.011. URL https://www.sciencedirect.com/science/article/pii/S1041608007000428.

Francesca Rossi, Kristen Brent Venable, and Toby Walsh. *A Short Introduction to Preferences: Between AI and Social Choice*. Morgan & Claypool Publishers, 1st edition, 2011. ISBN 1608455866. URL https://dl.acm.org/doi/10.5555/2049991.

Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. Verbosity bias in preference labeling by large language models. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*, 2023. URL https://openreview.net/forum?id=magEgFpK1y.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Alfred Schutz. *The Problem of Rationality in the Social World*, pp. 64–88. Springer Netherlands, Dordrecht, 1976. ISBN 978-94-010-1340-6. doi: 10.1007/978-94-010-1340-6_3. URL https://doi.org/10.1007/978-94-010-1340-6_3.

Marie Schäfer, Daniel B. M. Haun, and Michael Tomasello. Fair is not fair everywhere. *Psychological Science*, 26(8):1252–1260, 2015. doi: 10.1177/0956797615586188. URL https://doi.org/10.1177/0956797615586188. PMID: 26115962.

John H. Sears. Effective and lawful avoidance of taxes. *Virginia Law Review*, 8(2):77–85, 1921. ISSN 00426601. URL http://www.jstor.org/stable/1064452.

Alex Y. Seita. Uncertainty and contract law. *University of Pittsburgh Law Review*, 46(75), 1984. URL https://ssrn.com/abstract=1692858.

Amartya Sen. Social choice and justice: A review article. *Journal of Economic Literature*, 23 (December), 1985. URL https://scholar.harvard.edu/sen/publications/social-choice-and-justice-review-article. Review article on K.J. Arrow's Collected Papers: Social Choice and Justice.

Steven Shavell. Damage measures for breach of contract. *Bell Journal of Economics*, 11(2):466–490, 1980. URL https://EconPapers.repec.org/RePEc:rje:bellje:v:11:y:1980:i:autumn:p:466-490.

Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. Large language model alignment: A survey, 2023. URL https://arxiv.org/abs/2309.15025.

Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z. Ren, and Anirudha Majumdar. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions, 2024. URL https://arxiv.org/abs/2412.05563.

Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward hacking. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9781713871088. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/3d719fee332caa23d5038b8a90e81796-Abstract-Conference.html.

Irene Solaiman and Christy Dennison. Process for adapting language models to society (PALMS) with values-targeted datasets. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393. URL https://dl.acm.org/doi/abs/10.5555/3540261.3540709.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L. Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. Position: A roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=gQpBnRHwxM.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546. URL https://dl.acm.org/doi/abs/10.5555/3495724.3495977.

Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. Principle-driven self-alignment of language models from scratch with minimal human supervision. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/0764db1151b936aca59249e2c1386101-Abstract-Conference.html.

Kumar Tanmay, Aditi Khandelwal, Utkarsh Agarwal, and Monojit Choudhury. Probing the moral development of large language models through defining issues test, 2023. URL https://arxiv.org/abs/2309.13356.

Petros Terzis. Against digital constitutionalism. *European Law Open*, First View, July 2024. doi: 10.2139/ssrn.4896078. URL https://ssrn.com/abstract=4896078.

Jean Tirole. Incomplete contracts: Where do we stand? *Econometrica*, 67(4):741–781, 1999. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/2999457.

Alex Turner, Logan Smith, Rohin Shah, Andrew Critch, and Prasad Tadepalli. Optimal policies tend to seek power. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 23063–23074. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/c26820b8a4c1b3c2aa868d6d57e14a79-Paper.pdf.

Alexander Matt Turner, Dylan Hadfield-Menell, and Prasad Tadepalli. Conservative agency via attainable utility preservation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, pp. 385–391, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450371100. doi: 10.1145/3375627.3375851. URL https://doi.org/10.1145/3375627.3375851.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does LLM safety training fail? In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=jA235JGM09.

Marcus Williams, Micah Carroll, Adhyyan Narang, Constantin Weisser, Brendan Murphy, and Anca Dragan. On targeted manipulation and deception when optimizing LLMs for user feedback, 2024. URL https://arxiv.org/abs/2411.02306.

O.E. Williamson. *Markets and Hierarchies, Analysis and Antitrust Implications: A Study in the Economics of Internal Organization*. Study in the economics of internal organization. Free Press, 1973. URL http://www.jstor.org/stable/1817092.

Yasin Abbasi Yadkori, Ilja Kuzborskij, András György, and Csaba Szepesvári. To believe or not to believe your LLM, 2024. URL https://arxiv.org/abs/2406.02543.

Mohammad Yaghini, Patty Liu, Franziska Boenisch, and Nicolas Papernot. Learning to walk impartially on the pareto frontier of fairness, privacy, and utility. In *NeurIPS 2023 Workshop on Regulatable ML*, 2023. URL https://openreview.net/forum?id=R5MTSLPyYZ.

Haolan Zhan, Zhuang Li, Xiaoxi Kang, Tao Feng, Yuncheng Hua, Lizhen Qu, Yi Ying, Mei Rianto Chandra, Kelly Rosalin, Jureynolds Jureynolds, Suraj Sharma, Shilin Qu, Linhao Luo, Ingrid Zukerman, Lay-Ki Soon, Zhaleh Semnani Azad, and Reza Haf. RENOVI: A benchmark towards remediating norm violations in socio-cultural conversations. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 3104–3117, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.196. URL https://aclanthology.org/2024.findings-naacl.196/.

Simon Zhuang and Dylan Hadfield-Menell. Consequences of misaligned AI. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 15763–15773. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/b607ba543ad05417b8507ee86c54fcb7-Paper.pdf.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences, 2020. URL https://arxiv.org/abs/1909.08593.

Caleb Ziems, Jane Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. The moral integrity corpus: A benchmark for ethical dialogue systems. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3755–3773, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.261. URL https://aclanthology.org/2022.acl-long.261/.

Thomas P Zollo, Todd Morrill, Zhun Deng, Jake Snell, Toniann Pitassi, and Richard Zemel. Prompt risk control: A rigorous framework for responsible deployment of large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=5tGGWOijvq.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023. URL https://arxiv.org/abs/2307.15043.