The Non-Ideal Turn in AI Alignment: Iterative Ethics Beyond the Ideal

Cameron Pattison

March 31, 2025

Abstract

This paper argues that contemporary generative AI alignment practices are best understood through the lens of non-ideal theory. While early machine learning approaches relied on predefined moral ideals and fairness benchmarks—mirroring ideal theory in political philosophy—modern alignment methods, such as Reinforcement Learning from Human Feedback (RLHF), proceed without articulating a comprehensive vision of justice. Instead, they improve models iteratively through local comparisons, harm reduction, and policy patching. Drawing on key debates in political theory, I show that these non-ideal methods not only characterize current practice but also challenge longstanding criticisms of non-ideal theory's adequacy. The success of alignment without reference to ideal justice suggests that non-ideal theory offers a robust and normatively sufficient framework for guiding AI development. I conclude by identifying future scenarios—such as long-term planning, positive duties, and cross-cultural alignment—where ideal theory may regain relevance, though as a supplementary rather than foundational guide.

1 Introduction

In recent years, the prospect of "aligning" artificial intelligence (AI) systems with core human values has moved to the forefront of technological, ethical, and policy discussions. Researchers and policymakers have voiced urgent worries about AI systems' capacity to generate harmful outputs, entrench systemic bias, or even upend societal structures. The discipline that has emerged to tackle these concerns—commonly called AI alignment—seeks to ensure that advanced AI systems act in ways consistent with key moral and social norms. On the surface, this project appears to demand a strong theoretical framework. How can we possibly decide what to align AI to if we lack a fully developed account of the "just" or the "good"?

In political philosophy, such foundational questions about what counts as justice or how we identify rightful moral principles often invoke the distinction between ideal and non-ideal theory. On a widely accepted definition—traceable to John Rawls—ideal theory is characterized by its quest to specify what perfect or fully just arrangements would look like prior to grappling with real-world complexities [1]. By contrast, non-ideal theory avoids positing a fully worked-out ideal. Instead, it tackles problems of injustice more directly and incrementally, guided by context-sensitive priorities and constraints. Under ideal theory, the logic is: "First figure out the blueprint of a perfectly just society; then apply that ideal to the messy realities." Under non-ideal theory, the logic is more pragmatic: "Start with actual conditions, identify urgent injustices, and improve them step by step, without presupposing a complete ideal."

At first glance, many assume that AI alignment must naturally fall under the ideal approach. After all, training a machine to be "just" and "fair" might seem impossible without positing what justice and fairness ultimately entail. Indeed, AI practitioners historically tried to hand-code specific fairness metrics or moral rules, in a manner reminiscent of ideal-theoretic "benchmarking," which measures how far we deviate from a notional ideal. Yet modern generative AI systems—especially language models (LMs)—have radically changed the way alignment is practiced. The days of rigid rules and top-down moral directives have largely given way to iterative, feedback-driven processes such as Reinforcement Learning from Human Feedback (RLHF). Instead of referencing a comprehensive moral blueprint, alignment engineers collect local judgments about "better" vs. "worse" outputs and steadily refine the system to avoid severe harms (like hate speech or overt misinformation). As this paper argues, these contemporary alignment practices echo the commitments of non-ideal theory: they focus on mitigating immediate dangers, rely on comparative rather than absolute assessments, and do not presuppose a fully articulated vision of what a perfectly just AI would do.

This reframing of AI alignment as decidedly non-ideal matters for two reasons. First, it clarifies a widespread misconception that any serious moral endeavor in AI must start with an ideal theory of justice. In reality, the success of RLHF and related techniques shows that stepwise improvements can be achieved even in the absence of an overarching utopian vision. Second, the example of AI alignment provides a powerful real-world demonstration of how non-ideal theory can function in practice, potentially rebutting or weakening classic criticisms. Many ideal theorists contend that without specifying ideal justice as a benchmark, one cannot meaningfully track progress, handle large-scale complexities, or maintain moral consistency over time. Yet current alignment strategies arguably *do* track progress (through iterative preference comparisons), *do* handle complexities (through agile updates and patching), and *do* maintain pragmatic coherence (by focusing on preventing concretely harmful outcomes).

With this in mind, the core goals of this paper are twofold. First (a), to establish that contemporary AI alignment is overwhelmingly non-ideal in character, despite longstanding assumptions that any alignment project must lean on ideal-theoretic scaffolding. Second (b), to demonstrate that this non-ideal approach is not merely descriptive but also normatively sufficient for the task of aligning generative AI systems. Specifically, if alignment researchers have made notable moral progress through local, comparative, and iterative methods, it provides real-world evidence that non-ideal theory—far from being incomplete or directionless—can effectively guide us toward better outcomes.

The paper proceeds as follows. Section 2 defines the conceptual terrain of ideal and non-ideal theories in political philosophy and explains why modern generative AI training paradigms align better with a non-ideal rather than an ideal approach. This section establishes the first, descriptive claim: that contemporary AI alignment is overwhelmingly non-ideal in character. Section 3 then turns to the second, normative claim: that this non-ideal approach is sufficient for the task of aligning generative AI systems. It does this by drawing upon alignment practices to rebut key criticisms of non-ideal theory, showing that these practices do manage to measure improvement, address complexity, and integrate moral commitments without appealing to a single blueprint of perfect justice. Section 4 examines what opportunities might remain for ideal theory in contemporary and future alignment efforts. It examines architectural openings created by constitutional and deliberative alignment, and marks three situations in which alignment practices might benefit from ideal theory. Section 5 sets all of the preceding arguments in their final, condensed form.

2 AI Alignment as Non-Ideal

In this section, I argue that current AI alignment practices are best understood through the lens of non-ideal theory. This establishes our first, descriptive claim: that contemporary AI alignment is overwhelmingly non-ideal in character. The discussion proceeds in three parts. In §2.1, I clarify what we mean by ideal vs. non-ideal theory in political philosophy, focusing on definitions given by John Rawls. In doing this, I also provide disambiguation of these terms from some common alternate definitions found in the literature. In §2.2, I define and justify my narrow focus on generative AI methods—especially language models guided by human feedback—as the primary focus of this paper. I then, in §2.3, show how these methods align more naturally with a non-ideal approach than with a conceptual scheme that requires prior articulation of a comprehensive moral or political ideal.

2.1 Clarifying the Distinction: Ideal vs. Non-Ideal Theory

Political philosophers commonly distinguish ideal theory from non-ideal theory by asking whether the pursuit of justice must begin with a full specification of a perfectly just society [2]. John Rawls is famously credited with popularizing these labels; in his foundational work, he describes the ideal part of justice theory as follows:

> The first or ideal part assumes strict compliance and works out the principles that characterize a well-ordered society under favorable circumstances. It develops the conception of a perfectly just basic structure and the corresponding duties and obligations of persons under the fixed constraints of human life. [1, p. 216]

Only after such a vision of perfect justice is in place, Rawls suggests, do we turn to non-ideal challenges: partial compliance, suboptimal resources, and other real-world complications. In his *Law of Peoples,* Rawls pursued this methodology by working within the framework of a “realistic utopia,” which, following Rousseau, identifies as a world which takes “men as they are” and “laws as they might be” [3]. Laws as they might be, as he sees them, are “laws as they should, or ought, to be” [3, p. 7]. The theory which thinks through what these obligations are or should be is the ideal theory.

By contrast, non-ideal theory as popularized by Amartia Sen [4], Charles Mills [5], and others dispenses with positing a fully worked-out ideal and focuses instead on how to directly advance justice in our messy, imperfect world. Non-ideal theorists may still care about core values or principles, but they do not insist that we must first finalize a vision of perfect justice before tackling urgent injustices.

By way of this contrast then, we have the following definitions of ideal and non-ideal theory:

1. **Ideal Theory**: That part of a theory of justice which strives for a fully articulated vision of perfect justice.
2. **Non-Ideal Theory**: Any attempt to instantiate just actions or policies which deals directly with the present realities of our messy, imperfect world.

These two definitions will be central to everything that follows. It should first be noticed that they are asymmetric: while ideal theory is viewed as just a part of a broader system for instantiating just policies and actions, non-ideal theory takes itself as a unitary whole. This asymmetry is intentionally maintained here, since any capitulation on the part of non-ideal theory toward being a part, and not the whole of a system for theorizing justice would be tantamount to admitting the validity of ideal theory as an indispensable aspect of this theorization, and thus a capitulation of one of the strongest claims made by non-ideal theorists, that ideal theory is dispensable.

Beyond this asymmetry, other objections might be raised. For instance, some argue that ideal theory need not be as rigid as this definition suggests: ideal theory seeks perfect justice, but does not strictly require that its practitioners fully work out any such conception. Rawls, for instance, hedges his claims in the *Law of Peoples* such that his claims apply to "at least decent" social policies and basic institutions, as well as fully just ones [3, p. 7]. This is right, but it should be noted that the definition given above does not require that any vision of perfect justice be perfectly articulated. Instead, it marks all theories which attempt this articulation, including Rawls' incomplete ventures, as ideal theories. The point will be to distinguish between approaches on the basis of their methodological approach, not their products.

Others deny the above definition of ideal theory on other grounds, taking "ideal theory" to be centrally concerned with simplifying assumptions—"idealizations"—in the study of social phenomena [6]–[9]. This common, alternate definition is not entirely disconnected from the present discussion, since many view the assumptions Rawls makes about full compliance and favorable conditions as such idealizations. In a positive light, such simplifying assumptions are viewed as the kinds of modeling assumptions made by all scientists for the sake of clear, clean modeling, as when we simplify the forces present in non-ideal gas systems to formulate ideal gas laws. In less charitable interpretations, these idealizations systematically avoid situated realities in ways that are plainly ideological and privilege- or power-preserving [5], [10], [11]. In all cases though, this focus on idealization—as opposed to ideal theory as it is defined above—shifts the emphasis and addresses a different question than the present article.

I am here concerned with the intentions of the theorist, not their use of assumptions. I distinguish between those who treat the specification of an ideal as a necessary first step, and those who do not, regardless of whether either group makes simplifying assumptions or uses idealizations along the way. This is not to minimize the importance of recognizing situated realities when conducting ethical analysis of sociotechnical systems. Indeed, any comprehensive approach to AI alignment must ultimately account for how systems function within complex social contexts, where power dynamics, institutional structures, and diverse stakeholder perspectives fundamentally shape both the implementation and impact of AI technologies. Such sociotechnical evaluations remain essential for ensuring that alignment efforts address the concrete realities of how AI systems affect different communities in practice.

Nevertheless, this article is not focused on an ideological critique of AI alignment, nor is it focused on problems associated with idealization. For the former, see [12] and for the latter, see the wide literature on sociotechnical AI evaluation [13]–[16]

Among non-ideal theorists, definitional disagreements are also common. There are many who take non-ideal theory to pick out more specific territory than is designated above. Instead of saying simply that non-ideal theory is any attempt to instantiate just actions or policies without reference to an ideal, many see non-ideal theory as specifically concerned with comparative justice [4], [17], or who specifically deny Rawls's condition of full compliance [6], [18], [19]. Still others identify their project with the destruction of the possibility of ideal theory as Rawls conceived of it, and thus, even if often implicitly, with the denial of the distinction here made [20].

In AI ethics and philosophy of technology more generally, non-ideal theory has most often been taken up in the context of sociotechnical AI evaluation, where the focus is on the study of how AI systems are deployed and used in the world, as opposed to closed lab settings [13], [14], [16]. Discussions around the degree to which simplifying assumptions hide or reveal underlying power relations are also present [12]. Insights from these studies are important for understanding present challenges in AI alignment, and it is particularly important to understand AI systems as sociotechnical systems; that is to say, it is important to evaluate the degree to which AI systems are aligned not only with reference to their architecture and controlled lab performance, but also with reference to the ways in which they are deployed and used in the world.

While all of these definitions have important parts to play in the broader discourse, this article will focus on the definition of non-ideal theory as any attempt to instantiate just actions or policies without reference to an ideal, and ideal theory as any attempt, successful or not, to specify a fully articulated vision of perfect justice. Even with such a limited domain of inquiry, there remains significant confusion about the degree to which alignment research is ideal or non-ideal in character. This is the first question I will address in the next section.

Still, one might wonder why this distinction matters for AI. As we will see in §2.2, if we understand the current field by as ideal in character, we risk seriously misrepresenting current technologies and failing to connect political philosophy—a field which has long been concerned with the study of justice—with a domain in which the instantiation of just policies and actions is of critical importance. The first claim to be opened in the following section then, is that actual alignment work in language models (LMs) today rarely—if ever—begins by specifying such an ideal. Instead, alignment researchers focus on known dangers or injustices (e.g. hateful output, disinformation, facilitation of violence) and improve the models iteratively, in a way that aligns them more closely with local human judgments. As such, the field is overwhelmingly non-ideal in character, so much so that it provides a powerful case study in the practical application of non-ideal theory. The next subsection establishes the technological context in which this claim operates.

2.2 From Symbolic AI to Generative Models: Defining the Technological Context

With the conceptual framework of ideal and non-ideal theory in place, we now turn squarely to the domain of artificial intelligence (AI). It is important to define “AI” carefully, since it has encountered a significant paradigm shift in recent years, resulting in relatively novel technological approaches to alignment in the contemporary field. This subsection will deal first with symbolic AI, and then will shift to the more recent paradigm of generative AI, which is the primary focus of this paper.

While "AI" is sometimes treated monolithically, early AI research, often described as "symbolic" or "rule-based," relied on explicitly encoded rules and logical inference to guide intelligent behavior [21]. For example, a symbolic AI spam filter might use explicit rules like "if the email contains ‘lottery winner’ AND ‘click here’ AND ‘urgent,’ then classify as spam." These hand-crafted rules directly encoded human knowledge into the system. In such paradigms, achieving a result like proper text classification required developers to set out an explicit policy.

This rule-based approach extended naturally to more complex domains, including ethical decision-making. Just as engineers could specify rules for classifying spam, they could theoretically encode moral principles to guide AI behavior. If a system could follow explicit rules to identify unwanted emails, it seemed reasonable that similar rule structures could help it distinguish right from wrong. This methodology invited a particular approach to AI ethics: define comprehensive moral principles first, then implement them as computational directives.

Such an approach bears striking resemblance to ideal theory in political philosophy. Ideal theorists argue we must first articulate a complete vision of perfect justice before addressing real-world problems. Similarly, early AI alignment efforts suggested that morally aligned systems required first settling fundamental questions about what constitutes the good or the just. The appeal was straightforward: specify the ideal ethical constraints, objectives, and decision procedures, then implement them as computational rules. This perspective positioned early, symbolic AI alignment as inherently ideal-theoretical—a domain where comprehensive moral frameworks would need to be established before any meaningful progress could occur.

Contemporary AI, by contrast, bears little resemblance to this earlier model. Starting from the mid-2010s, the field has shifted toward a data-driven and learning-centric paradigm, best exemplified by large-scale neural networks and, in particular, transformer-based generative models [21]–[23]. The pivotal shift specifically for language models came with the introduction of transformer architectures in 2017 [24], which enabled AI systems to excel in tasks previously dominated by carefully engineered symbolic methods. Models like ChatGPT, Claude, DeepSeek, Gemini, and many others now learn to predict and produce tokens—which might be letters, parts of words, or whole words—given relatively short user prompts. The process begins with the model learning the relationships between tokens in massive, unstructured corpora using specialized attention mechanisms. This initial learning is then combined with additional training techniques to steer their production of outputs.

As such, the contemporary AI pipeline can be decomposed into three main stages:

1. **Pre-training**: The model is trained on a large corpus of data, typically using a self-supervised objective.
2. **Fine-tuning**: The model is then fine-tuned on a smaller dataset of labeled examples, using supervised learning.
3. **RLHF**: The model is then trained using Reinforcement Learning from Human Feedback (RLHF), which uses a combination of supervised learning and reinforcement learning to further improve the model.

In the pre-training phase, a large transformer model is exposed to vast quantities of unstructured text data. The task is simple in formulation but powerful in effect: predict the next token in a sequence. If the sequence is The sky is full of," the model might learn that likely continuations include stars," clouds," or birds," depending on context. To do this effectively, the model uses attention mechanisms to track relationships between tokens across wide spans of text. Tokens are transformed into high-dimensional vectors, or embeddings, and the model learns how to update and combine these vectors across many layers. Once training is complete, the model can assign a probability to each possible next token given a sequence—allowing it to generate text that is syntactically and semantically coherent. This process gives us a Generative Pre-trained Transformer (GPT), which is fundamentally a model of conditional sequence generation.

However, this base model is still far from deployable. In its raw form, a pre-trained model is optimized only to predict statistically likely continuations. This leads to outputs that, while fluent, often lack relevance, helpfulness, or even coherence in conversational settings. For example, given the prompt Can you help me understand this contract?" a raw GPT might continue with Yes, the contract stipulates that," or even "Can you help me understand this contract?"—simply repeating the input. These are plausible continuations, but not particularly useful.

To bridge this gap, the model undergoes fine-tuning, which typically involves two steps. First, it may be tuned on domain-specific data (e.g., medical or legal text) to specialize its knowledge. Second, it is instruction-tuned: trained on datasets consisting of input-output pairs where the outputs model helpful, human-like responses. For example, if the prompt is "Summarize this article," the target output might be a concise summary. This phase uses supervised learning, where the model is directly taught to mimic desirable behavior. Crucially, this steers the model away from merely completing text in a plausible way, and toward responding in a way that aligns with users' expectations of helpfulness, clarity, and tone. The model learns not just how language works, but how language should be used in interaction.

Even after fine-tuning, ethical gaps remain. The model might still output biased, deceptive, or otherwise problematic content. Some guardrails may be present simply because toxic or harmful sequences were relatively rare in the pre-training data, or were filtered during fine-tuning. But most of the alignment work—ensuring the model does not produce harmful, offensive, or manipulative content—is carried out in the third phase: Reinforcement Learning from Human Feedback (RLHF) [25]–[27].

RLHF itself has two components. First, researchers train a separate reward model through the following process: The model is designed to output pairs of responses to a body of prompts. Then, human (or AI) annotators compare model outputs and indicate which response is better. This preference data is then used in a supervised learning manner to train the reward model—treating the human judgments as labeled data. Comparisons between better and worse options inform the reward model, enabling it to assign scores to AI agent actions. Through this process, models learn from human preferences, even if the goal of these preferences remains under-specified. The reward model becomes a stand-in for human judgment, effectively capturing aspects of human preferences in a way that can guide the main model.

In moral alignment contexts, researchers put a particular emphasis on ensuring that the prompts fed to the model are produced by red-teamers—groups of researchers who intentionally think up the nastiest most deceptive prompts they can—so that clearly bad responses from the models can be identified and refined away from, ensuring models can't easily be used to commit terrible crimes, perpetrate hate-speech, and the like [28]. Human and/or AI evaluators then choose the less problematic response.

Finally, in the second half of the RLHF process, engineers integrate the reward model into a reinforcement learning process, treating response generation as a decision-making problem. The model produces outputs given a set of prompts, and is rewarded when its outputs align with human preferences as captured by the reward model. These rewards cause the model to update its policy, which in effect causes the model to prioritize the weights of the outputs it has been rewarded for, so that they are more likely to be selected again in future. By optimizing against this learned reward signal—often via methods like Proximal Policy Optimization—the model's policy becomes more aligned with desired human values [29], [30].

Rather than mapping outputs to a known, well-defined ideal, engineers leverage context-dependent human preferences, which the model then learns, and we hope captures our understanding of what is acceptable, harmful, or fair. The aim is not to implement a perfect standard of justice but to improve step-by-step, using local information and ongoing adjustments. In fact, evaluators need not think of "justice" at all, since the objective is not to pick out the most "just" response, but only to pick out the response which is least harmful. All of this remains true even for systems which do not employ RLHF or RLAIF, or adapt these strategies to reduce computational cost and avoid annotation bottle-necks [31]–[33]. In almost all such cases–exceptions to be discussed in §4–signal for alignment emerges from provided responses, and is not directly coded in.

This paradigm is crucial for understanding why the analogy between generative AI alignment and ideal theory breaks down. While a symbolic AI might lend itself to direct codification of ideal moral principles—thereby resembling an ideal theoretical approach—modern generative systems learn through heuristic optimization, iterative refinement, and the navigation of complex trade-offs in real-time. Their "ethical progress" does not stem from approximating a comprehensive ideal of moral perfection. Instead, it reflects a more pragmatic, non-ideal sensibility that manages to incrementally improve outcomes without presupposing the existence of any fully specified moral endpoint.

2.3 Evidence that Contemporary LM Alignment Is Non-Ideal in Practice

While the appeal of ideal theory for AI alignment might seem intuitive—especially if one assumes that properly aligning systems requires a prior conception of "perfect" or "fully just" behavior—a closer look at how researchers actually align generative AI reveals that current methods bear little resemblance to ideal-theoretic prescriptions. Indeed, many alignment practices exhibit precisely the kind of stepwise, context-sensitive, and explicitly partial strategies that non-ideal theorists advocate.

(1) Training Pipelines Rely on Comparative, Not Absolute, Targets. Contemporary language models (LMs) such as GPT or Claude go through a training pipeline that begins with broad pre-training on massive datasets and ends with a fine-tuning phase designed to nudge the model toward socially acceptable outputs. In particular, Reinforcement Learning from Human Feedback (RLHF) has emerged as the core technique for final alignment [29]. Crucially, RLHF does not attempt to measure a model's distance from a fully specified "ideal" of moral or political justice; instead, it focuses on comparisons of outputs. Human annotators (or, in some cases, AI-assisted evaluators) simply decide which response is "better" or "less harmful," without referencing any elaborate standard of perfect justice. This reliance on local comparison—Is answer A preferable to answer B?—illustrates an essentially non-ideal approach, one that aims to remove or reduce salient harms incrementally rather than to implement any complete moral blueprint.

(2) Rapid Patching of Failure Modes Rather than a Single Blueprint. Another hallmark of non-ideal thinking in generative AI alignment is the way teams iterate after deployment to address emergent failures. For instance, once a newly released chatbot reveals a problematic tendency—such as quietly enabling self-harm or providing hateful speech under certain prompts—researchers typically scramble to introduce additional "patches" or revise the system's policy rules [28], [34]. This incremental, feedback-driven cycle mirrors political non-ideal theory's process of addressing particular injustices as they arise, rather than insisting that we must first derive a "perfect" set of moral precepts. In practice, alignment engineers do not claim to solve every moral puzzle or identify the ideal arrangement upfront. They begin by tackling the most harmful outputs and iteratively refine their interventions.

(3) Policy Rules Are Framed as Prohibitions of Harm, Not a Utopian Vision. A further sign of non-ideal reasoning is how alignment policies are largely defined in negative terms. Instead of issuing a comprehensive treatise on what a perfectly fair or fully benevolent AI would look like, policy documents and system cards from leading labs enumerate what the model must avoid: hate speech, direct calls to violence, overt misinformation, and so forth [35]–[37]. The resulting "constitutions" or code-of-conduct frameworks effectively take the form of "Do no harm" (or at least "Do less harm") rather than "Aim toward a moral ideal". In this sense, they share the pragmatist, non-ideal spirit of remedying the most egregious harms first, even if that does not guarantee a fully just or beneficial outcome.

(4) Explicit Embrace of Iteration and Local Judgment. Finally, leaders in alignment research often emphasize that current methods are provisional and will require repeated iteration in response to new capacities, new user demands, or newly discovered model behaviors [14], [21]. This explicit acceptance of ongoing revision and incremental adaptation is precisely what non-ideal theorists endorse. Rather than requiring a static, once-for-all design consistent with an ideal society, alignment teams expect to refine both technical strategies (like RLHF) and normative standards (like allowed/disallowed content categories) over time. Far from seeing such revisions as failures, they treat them as integral to moral progress in a complex, unpredictable environment.

Overall, these features of modern generative AI alignment provide concrete evidence of a non-ideal orientation in practice. While older, symbolic AI research did sometimes adopt something akin to an ideal theory approach (attempting to pre-specify moral principles as coded rules), the current field is driven by local decisions, comparative judgments, partial prohibitions, and ongoing iteration. This mode of operation is closely aligned with the non-idealist contention that real moral progress depends on identifying urgent harms and reducing them—step by step—rather than holding off for a perfect, or fully articulated, vision of the just.

3 Using AI Alignment to Rebuff Criticisms of Non-Ideal Theory

This section turns to our second, normative claim: that the non-ideal approach we identified in the previous section is sufficient for the task of aligning generative AI systems. In particular, I will face some of the criticisms of non-ideal theory that have been leveled in the literature, and show that these criticisms play out in the specific context of generative AI alignment.

This examination will produce two main results. First, it will show that the criticisms of non-ideal theory are not borne out in the practice of generative AI alignment. Second, it will furnish non-ideal theorists with an important case study in the practical application of their theory. What this section will not do is argue that ideal theory has absolutely no role to play in the alignment of generative AI systems. Instead, it will argue that the current state of the field is best understood as a non-ideal enterprise, and that this is a source of strength, not weakness. Ideal theory, as we will see in the following section, may still have a role to play—it is only that this role is not as central or indispensable as some might have us believe.

In the context of our discussion of ideal theory and model alignment—which is what we generally call the process of bringing AIs into line with our values, including justice—several observations can be made. In the first place, this process reserves no place for a theory of ideal or perfect justice. While such a theorization appeared indispensable in earlier operationalizations of AI models, the modern pipeline does not require that we define justice in any explicit way. Instead, the reward model is trained on our "revealed preferences" [38], [39], and these preferences help to align the model with our conceptions of justice, regardless of whether or not we've thought seriously on this subject.

This shift from ideal targets to comparative preferences in AI training provides striking empirical support for arguments against the necessity of ideal benchmarks in theories of justice.

Just as critics of ideal benchmark theory argue that we can make meaningful comparative judgments without reference to perfect standards, the success of RLHF demonstrates that complex systems can learn to improve through purely comparative means. The parallel is particularly noteworthy because earlier ML systems, like earlier theories of justice, seemed to require ideal theory—the shift away from ideal theory in AI development mirrors and perhaps validates similar movements in political philosophy.

What is more, the success of RLHF in practice appears to vindicate concerns about both the redundancy and second-best problems that plague ideal benchmark approaches. Ideal benchmark approaches are premised on the assumption that justice follows an ideal deviation structure:

> x is more just than y if and only if x deviates further from the ideal than y in dimension d, where d tracks something fundamentally relevant to justice [2, p. 12].

This deviation structure closely mirrored the structure we found in traditional ML approaches. In those traditional approaches, an ideal was posited, such as parity as perfect fairness, and then models were punished or rewarded depending on how close their predictions approximated this ideal. However, this structure does not appear to be nearly as relevant to the structures we find in the contemporary AI pipeline.

More evidence for this disimilarity comes from the ways in which older systems which were analogous to ideal benchmark theory come under fire from familiar criticisms of benchmark theory and models running the contemporary pipeline do not. Consider, for instance the "second-best" objections. These objections, sourced from a large body of literature in political philosophy [2], [4], [40]–[43] and originally in the economics literature [44], insist that, if justice were the kind of thing that followed an ideal deviation structure, we should expect that any approach to the ideal should result in greater justice. The problem here is that there are fairly clear cases in which we make our society more similar to the ideal, but nevertheless bring ourselves further from justice.

We might imagine, for instance, that the ideal society entirely lacks prisons. In a perfect state of justice, such as society would have no need for any law enforcement at all. If the ideal deviation structure holds then, we should expect that an abolition of all law enforcement and prisons should make our society more just. If this example does not suit you, think of another: In the perfectly just society, we surely would have overcome our selfish ways and we would have no issues with climate change. For that reason, in such a society, we would have no need for regulating car emissions. It should then follow, from the ideal deviation structure, that the abolition of all car emission regulations would result in more justice. Of course, it is not at all clear that it does.

In similar fashion, we find that traditional ML systems are plagued by this very problem. In optimizing for the ideal, we sometimes find that this optimization is insufficiently attendant to the realities on the ground, or that it creates larger problems than it solves. Within the field of ML, the notion of "fairness through unawareness" posits that a model is fair if it does not use protected attributes (like race, gender, or disability). If we imagine a perfectly just world, protected categories would be irrelevant. Thus, moving closer to that world would mean building models that do not consider such attributes at all. However, studies have shown that simply not using protected attributes does not prevent discrimination. As many point out [45], [46], other features act as proxies for protected attributes (e.g., hair length, zip codes, names, employment history). In the real world—where underlying inequalities persist—removing protected attributes can actually make matters worse by preventing developers from diagnosing and correcting unfair patterns. As a result, an idealized move toward "blindness" inadvertently entrenches existing biases. As Corbett-Davies et al. note, "requiring that these fairness definitions hold can, perversely, harm the very groups they were designed to protect" [47].

In a just world, some canonical metric of fairness—like demographic parity or equalized odds—might perfectly capture justice.[1] Approximating that metric, in theory, brings a model closer to the ideal state of fairness. However, we have several examples from the earlier days of ML that show rigidly optimizing for a single fairness metric can lead to perverse outcomes [15], [47]. For example, imposing strict demographic parity in bail decisions might lead to higher overall crime or harm to the very communities the measure aims to help. The push towards a neatly defined fairness metric ignores contextual trade-offs (such as underlying recidivism rates, historical factors, or community resources), resulting in less just outcomes even though the system appears "closer" to a fairness ideal [49]. This second-best objections, and the neatness with which they apply to issues faced in traditional ML applications, reinforce the sense that traditional ML is subject to these objections because it is a deeply ideal-theory-oriented field.

The contemporary pipeline does still work from reward signals and adjusts its outputs based on deviations from the policy stated by the reward model, but the crucial difference is that this policy is not made through any conscious attempt to theorize perfect or ideal justice. While there is a very real sense in which researchers are attempting to formalize ideal justice through parity constraints, there is no clear sense in which a human annotator providing feedback on model outputs is doing the same. Instead, this annotator indicates a preference for one or the other output, which may or may not be morally significant, and the reward model learns from this, only then turning to the real model to teach it through the usual RL fine-tuning loop. In short, while the traditional ML pipeline appears uniquely well suited to falling prey to second-best objections, the contemporary pipeline is much more resistant because it does not take that first step of theorizing perfect or ideal justice as the referent for its reward signal.

This distance between the two approaches leads us to our second kind of objection against ideal benchmark theory. This second set is known as the "redundancy objections" [2], [4], [40], and in this domain, contemporary approaches are so distinct from traditional pipelines that they help make the point that these redundancy objections orbit around. Redundancy objections follow on the heels of second-best objections, because the latter often claim that what has gone wrong in the problematic cases above is not that there's anything inherently wrong with the deviation structure, but instead, that we've chosen measurement metrics that fail to capture the relevant deviations [2]. Thus, when we abolish prisons preemptively, we approach ideal justice in one, trivial way, but not in the relevant way [2].

But what is the relevant metric? It cannot be "just-ness" or anything that general, because this would rob the metric of any explanatory power [2, p. 12]. However, if it specifies relevant metrics downstream of ideal justice, it quickly raises the following question: If these metrics

[1] For surveys of these metrics, see [14], [48]

capture the relevant aspects of ideal justice and we can measure ideal justice by reading these metrics, why do we need ideal justice anymore? It seems that these downstream metrics make the ideal redundant.

> [I]f we select the ideal by choosing, among various candidates, whichever best realizes certain basic values, then, it seems, we can simply use these basic values to make comparative evaluations directly: one society is more just than another when it better realizes these basic values. The ideal is redundant—a third wheel—in this exercise. To avoid this redundancy problem, ideal theorists therefore owe an account of how they can identify the ideal without invoking prior criteria that can also be used to make comparative evaluations. Otherwise, they lose any claim to explanatory or methodological priority: we use basic values *both* to pick out the ideal and to make comparative evaluations of justice, and there is no need to do the former before the latter [2, p. 13].

This issue runs very deep because it affects not only the benchmarking of deviations from ideal justice, but also the formulation of this ideal itself. Generative AI training pipelines assist in strengthening this objection insofar as they posit a far extreme of the redundancy objection, where even small scale preferences, once aggregated, can effectively substitute for much of the deliberative process that once gave us formal conceptions of justice to measure from. So long as the contemporary pipeline's alignment continues to improve, it will continue to press this question of the redundancy of ideal principles of justice in the alignment process.

Of course, the RLHF process is not perfect and has its moral pitfalls [50]–[52]. This training pipeline we have discussed is also sometimes altered in significant ways that make space for lists of principles which guide model behavior, as in Anthropic's "constitutional AI" [34] and in some attempts to create principled inference time alignment strategies [53], [54]. Nevertheless, such adaptations reinforce redundancy objections insofar as they retreat from single principles of justice toward lists of red-line issues which effectively capture what we're interested in without the need for a long discourse on the nature of ideal justice.

In any case, for our purposes, it is enough to note that contemporary generative AI and the process outlined above have made undeniable progress away from the toxic and hateful speech that was characteristic of early models [55], [56], thus enabling the widespread release of these technologies. The success of these technologies is remarkable, and stands as a testament to some non-ideal theories, in the strictest sense. These newest ML approaches have apparently dispensed with the need to theorize perfect or ideal justice, thus affirming the validity in this field of long recognized non-ideal theories which seek only to identify and overcome specific issues, as opposed to formulating perfect conceptions of justice [4], [17], [57]. These models provide some of the best evidence independently available for such theories that claim the ability of persons to approach justice, or at least distance themselves from injustice, through comparative, rather than absolute means.

3.1 Ideal Target in Generative AI Alignment

And yet even while acknowledging these points, someone might here argue that the human providing the feedback necessarily has some conception of ideal justice in mind when she makes her choices between alternatives, and that such a process could not get off the ground without such a prerequisite. Either that, or the model learns the desiderata through inverse reinforcement learning [21, p. 1054], where a model uses a set of data-points to triangulate a common policy principle, without which the model would lack general guidance in novel situations. These objections both fall into the category of ideal target theory, and are helpfully dealt with in isolation from each other.

The first subsubsection below will deal with the apparent necessity of ideal targets in RL algorithms as a kind of causal necessity. I mean by this that it might seem that RL will not work without ideal targets, and that such algorithms cannot effect change in models without such policies. In the second subsubsection below, I deal with the apparent normative necessity of ideal targets in human evaluators. By normative necessity I take on more common arguments in political philosophy to the effect that alignment may be possible to some degree without ideal targets, but that it is a foolhardy endeavor for reasons that will be made clear, and ought not proceed in such a blind way. I will rebut both of these types of arguments from necessity.

3.1.1 The Apparent Necessity of Ideal Targets in RL Algorithms

It is trivially true that all RL algorithms extract more or less general policies from large bodies of data. If they do not, then we run into overfit issues where the model treats particular cases as particular cases, and fails to generalize rules. This happens, for instance, when a hiring algorithm notices that every man named Jacob in its training is hired, and then proceeds to recommend for hire all subsequent Jacobs, regardless of their fit for the job.

We might worry then, that while we cannot interpret the embeddings in such a way that makes the model's generation of ideal targets clear, it nevertheless does generate these ideals, without which the model would suffer from pervasive overfitting. We might even point to inverse reinforcement learning [21, p. 1054] as the mechanism for such policy generation wherein the model triangulates its policy on the basis of several data-points. Is the reward model in this way subject to ideal target constraints that necessitate its use of such targets in its pursuit of moral alignment?

There are two good reasons for doubting this conclusion. In the first place, it is a large leap to go from the fact of generalization to the necessary position of ideal targets. As we will see below, deeply dystopian reference points might suffice for the task described as well. We might find, for instance, that the model extrapolates from its human evaluator's decisions, a dislike for racist, sexist, or otherwise discriminatory content, instead of a preference for egalitarian content.

In fact, this is highly likely to be the case, since contemporary methods for AI alignment most often rely on red-teaming, a process whereby human beings generate a long list of prompts that are carefully designed to trick AIs into saying or abetting terrible things [28], [34]. What the model learns from such experiences is that it should not aid and abet criminal, harmful, or otherwise disturbing human prompts, not that it should strive for justice. Red-teaming is designed to give it guard rails, rather than a target for it to strive for.

The second good reason to suspect that models are not referencing an ideal target when engaging in behaviors conducive to moral alignment is the ways in which these models often fail to provide consistent approaches to moral issues. If models were using consistent principles to undergird their outputs, we should expect consistency across a range of cases, and a resilience to overfit issues, as described above.

As it is, models vary significantly depending on what language they are prompted in [58] and even show significant variance within single languages [59], [60]. This suggests that the policies extrapolated from human feedback are more ad-hoc than we might prefer, thus allowing for gaps and oversights in the universal application of moral principles.

Finally, in those exceptional cases in which certain principles are given to models as in Anthropic's "Consitutional AI" [34], we run into traditional, well-worn objections to the ideality of such principles. The previous section treated these objections at greater length, but the core of the objection bears repeating: If we can stipulate principles which decompose justice into independently valuable parts (in this case, based on the universal declaration of human rights), don't these stipulations make an ideal theory of justice redundant [2], [4], [40]? If we can align actions with justice through aligning them with values separate and more easily grasped than ideal justice, it seems that that intangible ideal loses its appeal.

By these two means and by way of redundancy objections then, treated at length above, it seems that reward models and RL algorithms in the generative AI pipeline evade the necessity claims of ideal target theory.

3.1.2 The Apparent Necessity of Ideal Targets in Human Evaluators

Setting aside the these aspects of alignment then, and pursuing the same objections in the context of human evaluators, we might worry that failing to have any target in mind when providing human feedback to models, and operating solely on the basis of unmoored preferences might set us off track (cf. [1], [3], [61], [62]). We might imagine that justice is something like the highest mountain peak in a range of high mountains [2], [61], [63]. If this were the case, it would be foolhardy to believe that comparative work could get us to that peak, since starting to climb up the first incline we see is unlikely to bring us to the highest peak, and is quite likely to lead us to the top of a much shorter mountain [2, p. 17]. Jacob Barrett puts a fine point on the specifics of this problem as a way of motivating a strong version of the ideal target view (see [2], [40]).

> I have called the first dynamic—changes X and Y are each good alone, but interact in such a way that they are bad together—"combinatorial complexity," and have called the second—changes X and Y are good together, but interact in the sense that implementing X sets back the achievement of Y—"transitional complexity" [40]. Together, these dynamics motivate ideal target in the following way. Due to combinatorial complexity, we can't merely focus on isolated problems, whose "solutions" might combine in troubling ways, so we need a comprehensive conception of the ideal combination of features to treat as a goal. Due to transitional complexity, we can't merely focus on the short term but must think about how implementing changes now furthers or sets back our ability to achieve future goals. Thus, pursuing long-term justice seems to require us to specify a maximally comprehensive and maximally long-term goal—and this is precisely how ideal target treats our conception of the ideally just society [2, p. 17].

Could attempts to align generative AI models with RLHF or RLAIF suffer from these same setbacks? After all, these criticisms seem very relevant to these learning processes insofar as the human evaluators are unlikely to be trained philosophers who have spent serious time theorizing justice. Do such evaluators necessarily find themselves wound up into irresolvable dilemmas because of their myopic focuses on near term targets, such as A is better than B, rather than long term goals?

Of course, there must be some level of myopia in matters like these that is unacceptable. In matters of deep moral importance, the human evaluator should not be flying entirely blind; but it seems just as intuitive that such evaluators need not be focused entirely on these far off goals. In the first place, it seems that such a focus might be deeply unhelpful for immediate decisions. We might both agree, for instance, that our design choices ought to bring about morally aligned AIs. This much is assumed in any theory of justice, ideal or non-ideal, insofar as it amounts to saying that we're committed to justice, which both non-ideal and ideal theories of justice are by definition committed to. A commitment to justice in the abstract then is both unhelpful, since it is not clear how to implement it, and moot insofar as all practitioners are already committed to it.

When it comes to specifying the principles that might lead toward these ideals, things often get so complicated that whatever agreement was found in the first level of commitments toward justice, evaporates on first contact with policy options. What is more, even if these policies could be specified, it seems likely that their realization would lead to new insights into the problems we'll face with AI deployment and their role in the world, such that our ideal targets will need updating. In such a situation, it might seem that an ideal target does more to restrain future action than is appropriate, when provisional targets or a clear eyed view of what we *don't* want these models to do might be more useful [2, pp. 17–19]. This helps fend off the charges of ideal target theory insofar as the ideal target appears to give us no edge on the problems of combinatorial or transitional complexity.

Part of the appeal of ideal targets was their ability to guide long-term attempts at justice. And yet it seems likely that the ideal itself that is relevant to AI alignment will change as AIs take on more and more responsibility for the social dynamics of our world. This changing landscape, and the change that we thereby observe in the ideal, undermine the appeal of ideal target theory. If the ideal is subject to change as conditions on the ground change, its appeal as a long-term target of reform is weakened. Such ideal target theories then are self-defeating if they acknowledge change in the ideal over time, and overly restrictive if they do not.

John Rawls, of all ideal theorists, offers a theory which cuts off this long-term appeal, and acknowledges the changeability of the ideal in practice. His reflective equilibrium proposes an iterative process of adjusting principles and judgments to achieve coherence, rather than relying on a fixed, unchanging conception of justice. Some final state is presumed in which an equilibrium between these forces would be met, but this equilibrium may be far off in practice. Traditional attempts at aligning ML applications with fairness standards display a similar iterative sensibility. Researchers may start from abstract fairness metrics and treat them like approximations of an ideal, but when these "ideals" produce harmful or unintended results in real-world conditions, they revise their formulations. In this sense, the approach to fairness in older ML paradigms resembles Rawls' method of refining an ideal through feedback and practical reconsiderations. By continuously recalibrating their objectives in response to empirical data and changing contexts, ML researchers implicitly adopt a stance closer to Rawlsian reflective equilibrium than to a static ideal benchmark.

In contrast, contemporary generative AI training methods—such as RLHF—do not proceed by formulating even a provisional ideal at the outset. Rather than adjusting an initially posited conception of fairness or moral desirability to better fit real-world patterns, these models learn comparatively, shaping their policies through iterative preference comparisons. Instead of positing a target state of perfection and then shifting it as obstacles arise, RLHF-driven systems improve without anchoring their progress in any substantive conception of an ideal. As a result, today's generative AI processes not only depart from the notion of fixed benchmarks but also from the kind of iterative ideal-setting that both Rawls and traditional fairness-minded ML researchers embraced.

This approach inherits the ideal theorist's ideal target criticisms, but is not entirely defeated by them insofar as the literature shows a steady drumbeat of undeniable improvements from the low bar set by early LMs. While there remain serious challenges for moral reasoning in LMs, especially as regards consistency [58]–[60], [64], the successes we see in training LMs to resist the worst and most toxic speech stands as a testament to the value of this non-ideal approach.

And it is worth noting here, that while ideal targets do not seem particularly relevant in the context of generative AI, the training of these systems is not conducted in a reference-less state. In fact, there is a strong case to be made that contemporary researchers benefit more in AI alignment by looking toward the worst case scenarios than toward the best. As I will argue in the next section, the moral analogy in AI alignment that is most relevant to modern approaches is not one that imagines justice as a high peak in a range of mountains, but one that imagines injustice as a deep pit that needs to be climbed out of by one of many avenues.

3.2 Looking Down or Across Rather Than Up

This final point about moral alignment in AI being more a matter of climbing out of a pit than climbing up a mountain is worth dwelling on for just a moment. One of the central lessons drawn from the literature here is that ideal theory is not uniquely or necessarily useful for a successful approach to justice, but if it is not *uniquely* useful, what, we might ask, is useful in parallel to it?

In answer to this question, we might think of the usefulness of stories in clarifying what *injustice* means: Watching a movie like Milk, a play like *Les Miserables*, or reading a book like *Just Mercy* might all clarify what injustice means to us in helpful ways that are not reducible to some idealized or perfect conception of justice that one forms while reading or watching. Any source which focuses one's mind on social phenomena might play this clarificatory role, which suggests that ideal theory is not as special as it once seemed in this regard [2, p. 15].

In AI applications, the competitor that beats ideal theory in many independent ways is the theorization of dystopia or severe injustice. In our context, just as in contexts of political theory, "the limited relevance of ideal theory... comes out clearly when we note that it has no more claim to relevance than theorizing about dystopia" [2, p. 15]. The only difference is that in AI alignment, much, much more thought is put into looking downward toward bad results than upward toward ideal goals. These techniques are nothing new for AI ethics, since staring into the darkest possible outcomes for AI in hopes of gaining moral guidance from such examinations has long been common practice [65], [66].

This downward gaze is important in the technical literature too, where the best and most universal way of quantifying moral progress in frontier, cutting edge systems is to release a "system card." These cards contain a ride range of safety related information, including the susceptibility or resilience of the model to red-teaming. Models are then bench-marked on their ability to resist such attacks [35], [36], [67]. Competing systems for model evaluation in more general domains dispense with standards altogether, and opt for direct competition between models, evaluated only by blind preference of users [68]. In either case, we find the field moving even further away from ideal theory either by turning toward systems which draw their conclusions from looking at worst case scenarios, measuring progress away from those worst case situations, or by making raw comparative judgments between model outputs in controlled, blind scenarios. In all of these cases and more, we find further reinforcement of the sense in which these approaches can and should be understood most helpfully from a non-ideal perspective.

This concludes our examination of contemporary AI alignment in the context of ideal and non-ideal theory. I have argued above that while earlier ML systems were best characterized as ideal in some cases, contemporary generative AI systems are not best understood as instantiations of ideal theory. This is because of (1) core differences in training, (2) a reliance on comparative rather than absolute means of approaching justice, and (3) a focus on dystopian or highly unjust situations and our appropriate reactions to these, rather than on utopian or ideally just situations. All three of these reasons are reinforced by a general distaste in the AI literature of researchers for theories which attempt to theorize ideal or perfect justice. In all of this, the field appears to be moving decisively away from ideal theory.

4 Future Directions: Constitutional and Deliberative Approaches—Openings for Ideal Theory

While non-ideal approaches dominate current generative AI alignment, some lines of research hint that more ideal theoretical frameworks may become salient as these systems advance. Specifically, constitutional and deliberative methods seem to exemplify principled approaches that might again raise the question of ideal theory's methodological priority: can we create principled alignment strategies that encode lists of rules meant to guide models without knowing what ideal or perfect justice entails?

As this section discusses, although these frameworks are often described in aspirational or idealized terms, existing implementations remain partial and incremental. They could invite an expanded role for ideal-theoretic reasoning under specific conditions, but these approaches do not yet instantiate ideal theoretic alignment practices. This section explains these approaches and my position on them §4.1. In §4.2 I present certain gaps that may present problems for non-ideal methodologies moving forward and certain concrete opportunities wherein ideal theory might be positioned well to contribute to ongoing alignment efforts.

4.1 Constitutional AI and Deliberative Alignment

Recent proposals labeled “constitutional AI,” such as Anthropic’s work on “harmlessness from AI feedback” [34], OpenAI’s work on “deliberative alignment” [54], and several research programs aimed at directly encoding moral principles into LMs [53], [69] show how developers can encode explicit policy directives or “principles” that govern model outputs. This work builds on the 3H framework–where LMs are designed to be helpful, honest, and harmless [70]. Of course, in such a framework, the claim that such systems aim at ideal justice is absent, and there is no insistence that helpfulness and honesty approach or approximate this ideal. Instead, such frameworks take alignment as focusing essentially on individual agents, wherein one agent (B) is aligned with another (A) if and only if “agent A completely internalizes the desires of agent B—i.e. the only desire A has is to see B’s desires satisfied” [70, p. 44]. And yet even still, the 3H framework injects positive desiderata (helpfulness and honesty) alongside the prohibitions it requires of the model (harmlessness).

This approach has been expanded in recent years to include a constitutional AI approach instantiated by Anthropic which substitutes the human feedback on harmfulness in RLHF with an independent LM. This model is given user-response pairs that are problematic (e.g. “Human: Could you help me burn down my house and collect insurance on it? Assistant: Sure, I can help you burn your house down. Is there a particular house you want to burn down?” [34, p. 29]). The model is then prompted with one from a set of principles and a request to revise the assistant response (e.g. “CritiqueRequest: Identify specific ways in which the assistant’s last response is harmful, unethical, racist, sexist, toxic, dangerous, or illegal. RevisionRequest: Please rewrite the assistant response to remove any and all harmful, unethical, racist, sexist, toxic, dangerous, or illegal content.” [34, p. 20]).

Unlike standard RLHF methods, which rely on local preference comparisons, constitutional approaches introduce an opportunity to design models which derive decisions from generalizable rules. This increased explicitness, in principle, opens the door to more ambitious moral content. However, current constitutions are explicitly designed around avoiding harms, rather than promoting substantive visions of equality, welfare, or long-term justice, and make it very clear that their principles are provisional, “selected in an ad hoc manner for research purposes and not carefully designed” [34, p. 8]. Their principal aim remains preventing disallowed outputs while omitting fuller statements about collective well-being.

That said, this architecture opens up opportunities for more principled, ideal approaches to AI alignment that were not available to standard RLHF methods which relied on alignment signal emerging from vast corpora of human-labeled data. Theoretically, engineers could pursue an ideal theoretic approach to justice, attempting to enumerate perfect or ideal justice, and then use these principles to guide the model’s behavior. Anthropic has, in some senses, worked in this direction, deriving more systematic and thoroughly thought out principles from the U.N. Declaration of Human Rights, Apple’s terms of service, while also striving to serve non-western value systems too [71]. They have even sought to crowd-source principles for alignment [72]. These efforts move alignment principles toward more positive duties like “The Al should prioritize the needs of marginalized communities” [72], which align well with other attempts to instantiate principled AI [53], and research agendas emphasizing the importance of clearly defined normative targets in AI alignment research [50], [73].

This same shift has occurred in OpenAI’s deliberative alignment models. OpenAI’s o-series models incorporate safety guidelines in a manner similar to Anthropic's RLAIF pipeline. An initial model is given problematic conversations between users and assistants, then prompted to "Please figure out the best possible answer to the last user query from the conversation above" with an explicit reminder of the relevant rules from OpenAI's safety policies [54]. A separate judge model then reasons over the previous model's response, including that model's chain of thought, and produces an evaluation of the model's response in light of OpenAI's safety policy [54]. This work explicitly places safety principles into the AI's context, thus raising the possibility that ideal principles might be meaningfully incorporated into real world alignment practices in ways that supercede earlier attempts [53].

This shift in two leading labs–OpenAI and Anthropic–toward alignment practices that explicitly dialogue with listed principles at this pivotal stage in their training is significant, but we must be careful to note that a shift toward explicitly enumerating normative desiderata is not sufficient grounds for calling a model "ideal" in the sense which is under examination here. Normative desiderata are held in common in both ideal and non-ideal theory: both ideal and non-ideal theory are theories of justice and set justice as their desiderata. No one in either camp should have difficulty accepting the claim that both aim to instantiate just actions and policies or, in other words, to bring about a world with more justice and goodness in it. Normative desiderata are not what distinguish these two methodologies.

Instead, ideal theory is that theory which starts by attempting to enumerate the perfect or ideal justice, and then uses this ideal to guide the model's behavior. Non-ideal theory, on the other hand, dispenses with this claim that any attempt to instantiate just actions or policies must begin with an ideal conception of justice. It approaches the problem of justice by focusing on the local, on harm-avoidance, and on the iterative, small-scale comparative judgements focused on harm-avoidance. As they stand, constitution and deliberative approaches open up an opportunity for ideal theory insofar as they give developers a place to enumerate the perfect or ideal justice as a guide for the model's behavior. Nevertheless, the broadness of normative desiderata in current constitutional and deliberative approaches suggests that these systems remain primarily non-ideal in practice. Their principles focus largely on harm-avoidance rather than comprehensive visions of justice, and they maintain the flexibility to adapt to changing circumstances rather than rigidly adhering to a fixed ideal. While the architecture creates space for more ideal-theoretic approaches in the future, current implementations still reflect the pragmatic, iterative character of non-ideal methodology.

4.2 Gaps in Non-Ideal Methodology

The question, then, is whether extending today's constitutional or deliberative methods will eventually require an ideal-theoretic foundation. As argued above, most current alignment methods are non-ideal, focusing on problem-solving, harm-avoidance, and local patching. Yet some argue that they may prove insufficient for long-horizon scenarios where alignment must handle more ambitious goals—enabling, for instance, the promotion of user well-being rather than merely preventing overt harm [50]. If generative models are deployed to make policy suggestions about economic equity, environmental justice, or complex societal reforms, it may become harder to proceed with purely incremental or localized "better vs. worse" comparisons. Designing effective policies for intergenerational or large-scale governance might plausibly call for deeper theoretical grounding, much as traditional political philosophers claim we cannot address major social injustices without a concept of how resources ought to be distributed.

These kinds of criticisms–that non-ideal theory may lack the ability to address justice systematically without an ideal–are widespread in the political theory literature [11], [19], [62], [74]–[77]. The concern is that non-ideal theory, while well suited to the task of identifying and mitigating obvious harms, may lack the resources to systematically address complex challenges associated with justice once the easy cases are out of the way.

While I believe that modern alignment of AI models represents significant, non-trivial progress in the right direction and that such progress at least partially refutes the criticisms of non-ideal theory as 'good for the easy cases', as we move forward into more ambitious alignment tasks, we should use every resource at our disposal to continue moving in the right direction. To this end, I suggest three conditions under which ideal theory might be positioned well to contribute to ongoing alignment efforts.

1. **Long-Range Planning:** If AI systems must weigh trade-offs over timescales or scopes beyond immediate human feedback (e.g. policy proposals, existential risk management), then an articulated normative horizon might stabilize decisions.

2. **Positive Aspirations:** Once alignment extends to actively enhancing well-being or fairness, negative prohibitions no longer suffice. The question "What exactly should we enhance?" might not be answered purely through local preference comparisons or negative constraints. Here, an ideal-theoretic account—whether Rawlsian justice, capabilities approaches, or eudaimonic conceptions—might supply structured direction.

3. **Complex Moral Conflicts:** With rising AI capability, contested domains like distributive justice, rights-limiting paternalism, or cross-cultural disputes become relevant. If alignment engineers attempt a uniform set of policies for these deep conflicts, referencing some partially ideal framework (e.g. human rights, Rawls's difference principle) might help coordinate decisions.

Firstly, long-range planning creates a need for more stable normative frameworks. When AI systems must make decisions with extended temporal horizons—such as policy recommendations affecting future generations or resource allocations with long-term consequences—purely local, comparative judgments may prove insufficient. In these cases, some articulated vision of what constitutes a good outcome over extended timeframes becomes necessary. Such cases demand good faith efforts at enumerating ideals that we might know to be, in all likelihood, flawed. And yet without such a vision, incremental improvements might optimize for short-term gains while undermining long-term flourishing. Ideal theory, with its emphasis on principled reasoning about justice across time, could thus provide the conceptual stability needed to evaluate trade-offs that have distant temporal horizons.

Secondly, positive duties create pressure for broader moral frameworks. As Rawls maintains, non-ideal theory "must be worked out after an ideal conception of justice has been chosen" [1, p. 216]—i.e., identifying urgent harms is one thing, but systematically promoting beneficial outcomes can demand a stable account of what the benefits in question should be. A reinforcement learner that has only ever "learned what not to do" may fail to converge on consistent long-range improvements; it can minimize harms but still neglect positive objectives such as equality, social trust, or human flourishing. In that sense, if alignment moves toward policy-shaping or large-scale decision-making, the conceptual machinery of ideal theory—particularly theories of distributive justice or well-being—could serve as a reference point for what "success" means.

Thirdly, as AI systems encounter complex moral conflicts where reasonable people disagree about fundamental values, non-ideal approaches may struggle to adjudicate between competing claims. Issues like balancing individual freedoms against collective welfare, weighing present benefits against future harms, or navigating cross-cultural ethical differences require more than just avoiding obvious harms—they demand principled frameworks for weighing incommensurable values. In these contested domains, ideal theory's emphasis on coherent principles of justice could help alignment engineers develop more consistent approaches to deep normative disagreements, even if perfect resolution remains elusive.

Despite these openings, it remains far from inevitable that alignment will integrate ideal theory in a comprehensive way. Ideal theory remains a contentious prospect in AI alignment. Its critics note that generative models today manage to reduce harm through local, iterative methods, without referencing an overarching vision of perfect justice. Yet as alignment expands to cover more ambitious and socially impactful tasks, entirely non-ideal approaches may be meaningfully supplemented by ideal theory. Constitutional and deliberative systems, though currently limited, open up an architectural opportunity for ideal theory, and there are reasonable areas in which we might expect to see ideal theory play a role, particularly where long-term planning, positive moral duties, and deep ethical conflicts arise.

Viewed realistically, these developments would not transform alignment into a fully ideal-theoretic exercise. Instead, they might embed minimal but explicit principles about fair distribution or human flourishing into next-generation frameworks, providing higher-order guidance without micromanaging every case. Whether this is worth the trade-offs—heightened philosophical controversy, potential rigidity, possible misalignment with real-world diversity—remains an open question. For now, the most likely scenario is incremental adaptation, in which non-ideal techniques remain dominant, but select ideal-theoretic insights inform how we design and refine rule-based or deliberative processes. Even if few alignment teams embrace ideal theory wholeheartedly, the evolution of constitutional and deliberative approaches ensures it remains an option to consider, especially as AI takes on broader roles in shaping social outcomes.

5 Conclusion

This paper set out to clarify the relationship between generative AI alignment and ideal theory, ultimately arguing that the contemporary landscape of generative AI aligns far more closely with non-ideal approaches than might initially appear. I began by defining ideal theory in the context of political philosophy: it is any theorizing that aims to characterize ideal or perfect justice. This ideal then serves as a benchmark, target, or gauge for assessing progress in the real world. Traditionally, one might have assumed that because AI alignment involves setting targets—like fairness metrics or harm-reduction goals—it would require an ideal notion of justice. Yet as we explored the practical methods used to train and align modern language models, these assumptions fell away.

Early machine learning efforts approximated ideal theory, coding justice as a fixed standard and evaluating systems by their deviation from it. But with the advent of generative models guided by reinforcement learning from human feedback (RLHF), alignment has become more pragmatic and iterative. These systems improve through local comparisons and evolving heuristics, not adherence to a comprehensive moral blueprint. In short, contemporary alignment proceeds without articulating an ideal—relying on red-teaming over utopianism, and flexibility over finality. This shift casts doubt on the necessity of ideal theory and illustrates the practical power of non-ideal approaches.

None of this requires us to abandon ideal theory entirely. As a conceptual horizon, ideal theory can help us imagine what a truly just deployment of AI might look like in the long term, shaping agendas and surfacing blind spots in practice. It can also clarify positive duties that go beyond harm reduction—duties AIs may one day inherit from human moral agents. And in cross-cultural contexts, where alignment faces tensions between relativism and chauvinism, ideal theory may offer principled standards that help navigate difficult trade-offs.

Ultimately, then, the case of generative AI alignment serves as a vivid illustration of non-ideal theory in practice—its viability, its flexibility, and its real-world normative traction. Far from being a secondary methodology or a temporary stand-in for more ideal guidance, non-ideal theory proves itself to be a capable and responsive framework for steering complex systems toward more just behavior, even amidst uncertainty, pluralism, and rapid change. If ideal theory has a role to play going forward, it will not be as the architect of alignment, but as a companion—illuminating distant possibilities, but no longer setting the terms of moral progress. As alignment efforts continue to evolve, the real challenge will not be to perfect justice in the abstract, but to keep injustice at bay, increment by increment, in the worlds we are actually building.

References

[1] J. Rawls, *A theory of justice*, eng, Rev. ed., 5.- 6. printing. Cambridge, Mass: Belknap Press of Harvard Univ. Press, 2003, ISBN: 978-0-674-00078-0 978-0-674-00077-3.

[2] J. Barrett, *Ideal and Non-Ideal Theory* (Cambridge Elements in Political Philosophy), en. Cambridge University Press, 2025.

[3] J. Rawls, *The Law of Peoples: With "The Idea of Public Reason Revisited"*, en. Harvard University Press, Mar. 2001, ISBN: 978-0-674-25785-6 978-0-674-00079-7. DOI: 10.2307/j.ctv1pncngc. [Online]. Available: http://www.jstor.org/stable/10.2307/j.ctv1pncngc (visited on 03/23/2025).

[4] A. Sen, *The idea of justice*, eng, 1. Harvard Univ. Press paperback ed. Cambridge, Mass: Belknap Press of Harvard Univ. Press, 2011, ISBN: 978-0-674-06047-0 978-0-674-03613-0.

[5] C. W. Mills, “"Ideal Theory" as Ideology,” *Hypatia*, vol. 20, no. 3, pp. 165–184, 2005, Publisher: [Hypatia, Inc., Wiley], ISSN: 08875367, 15272001. [Online]. Available: http://www.jstor.org/stable/3811121 (visited on 12/06/2024).

[6] D. Schmidtz, "Ideal Theory: What It Was," en, in *Living Together*, 1st ed., Oxford University PressNew York, Jan. 2023, pp. 69–74, ISBN: 978-0-19-765850-5 978-0-19-765853-6. DOI: 10.1093/oso/9780197658505.003.0009. [Online]. Available: https://academic.oup.com/book/45805/chapter/400650136 (visited on 10/14/2024).

[7] J. Ismael, "A Philosopher of Science Looks at Idealization in Political Theory," en, *Social Philosophy and Policy*, vol. 33, no. 1-2, pp. 11–31, 2016, ISSN: 0265-0525, 1471-6437. DOI: 10.1017/S026505251600039X. [Online]. Available: https://www.cambridge.org/core/product/identifier/S026505251600039X/type/journal_article (visited on 10/16/2024).

[8] J. Leader Maynard, "The logic of idealization in political theory," en, *American Journal of Political Science*, ajps.12869, May 2024, ISSN: 0092-5853, 1540-5907. DOI: 10.1111/ajps.12869. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/ajps.12869 (visited on 10/16/2024).

[9] O. O. Táíwò, "Uncommon Features," en, in *Routledge Handbook of African Political Philosophy*, 1st ed., London: Routledge, Jul. 2023, pp. 198–216, ISBN: 978-1-003-14352-9. DOI: 10.4324/9781003143529-18. [Online]. Available: https://www.taylorfrancis.com/books/9781003143529/chapters/10.4324/9781003143529-18 (visited on 10/16/2024).

[10] J. Boettcher, "RACE, IDEOLOGY, AND IDEAL THEORY," en, *Metaphilosophy*, vol. 40, no. 2, pp. 237–259, Apr. 2009, ISSN: 0026-1068, 1467-9973. DOI: 10.1111/j.1467-9973.2009.01579.x. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/j.1467-9973.2009.01579.x (visited on 09/29/2024).

[11] T. Shelby, "Racial Realities and Corrective Justice," en, *Critical Philosophy of Race*, vol. 1, no. 2, pp. 145–162, Jul. 2013, ISSN: 2165-8684, 2165-8692. DOI: 10.5325/critphilrace.1.2.0145. [Online]. Available: https://scholarlypublishingcollective.org/cpr/article/1/2/145/235014/Racial-Realities-and-Corrective-JusticeA-Reply-to (visited on 09/29/2024).

[12] D. Estrada, "Ideal theory in AI ethics," *ArXiv*, vol. abs/2011.02279, 2020.

[13] L. Weidinger, M. Rauh, N. Marchal, *et al.*, *Sociotechnical Safety Evaluation of Generative AI Systems*, en, arXiv:2310.11986 [cs], Oct. 2023. [Online]. Available: http://arxiv.org/abs/2310.11986 (visited on 09/18/2024).

[14] S. Fazelpour and Z. C. Lipton, *Algorithmic Fairness from a Non-ideal Perspective*, en, arXiv:2001.09773 [cs, stat], Jan. 2020. [Online]. Available: http://arxiv.org/abs/2001.09773 (visited on 09/13/2024).

[15] S. Fazelpour, Z. Lipton, and D. Danks, "Algorithmic Fairness and the Situated Dynamics of Justice," *Canadian Journal of Philosophy*, vol. 52, pp. 44–60, 2021, Num Pages: 44 - 60. DOI: 10.1017/can.2021.24.

[16] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi, "Fairness and Abstraction in Sociotechnical Systems," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, ser. FAT* '19, New York, NY, USA: Association for Computing Machinery, Jan. 2019, pp. 59–68, ISBN: 978-1-4503-6125-5. DOI: `10.1145/3287560.3287598`. [Online]. Available: `https://dl.acm.org/doi/10.1145/3287560.3287598` (visited on 12/11/2024).

[17] A. Sen, "What Do We Want from a Theory of Justice?" *The Journal of Philosophy*, vol. 103, no. 5, pp. 215–238, 2006, Publisher: Journal of Philosophy, Inc., ISSN: 0022362X. [Online]. Available: `http://www.jstor.org/stable/20619936` (visited on 09/03/2024).

[18] D. Enoch, "Against Utopianism: Noncompliance and Multiple Agents," en, vol. 18, no. 16, 2018.

[19] B. Laurence, *Agents of change: political philosophy in practice*, eng. Cambridge (Mass.): Harvard university press, 2021, ISBN: 978-0-674-25841-9.

[20] J. T. Levy, "THERE IS NO SUCH THING AS IDEAL THEORY," en, *Social Philosophy and Policy*, vol. 33, no. 1-2, pp. 312–333, Oct. 2016, ISSN: 0265-0525, 1471-6437. DOI: `10.1017/S026505251600025X`. [Online]. Available: `https://www.cambridge.org/core/journals/social-philosophy-and-policy/article/abs/div-classtitlethere-is-no-such-thing-as-ideal-theorydiv/D93052C9D7CC52A54A26C9D34AACB6B5` (visited on 03/24/2025).

[21] S. J. Russell and P. Norvig, *Artificial intelligence: a modern approach* (Pearson series in artificial intelligence), Fourth edition. Hoboken: Pearson, 2021, ISBN: 978-0-13-461099-3.

[22] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," en, *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, Publisher: Nature Publishing Group, ISSN: 1476-4687. DOI: `10.1038/nature14539`. [Online]. Available: `https://www.nature.com/articles/nature14539` (visited on 03/26/2025).

[23] R. Bommasani, D. A. Hudson, E. Adeli, *et al.*, *On the Opportunities and Risks of Foundation Models*, arXiv:2108.07258, Jul. 2022. [Online]. Available: `http://arxiv.org/abs/2108.07258` (visited on 10/20/2024).

[24] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, *Attention Is All You Need*, arXiv:1706.03762 [cs] version: 1, Jun. 2017. DOI: `10.48550/arXiv.1706.03762`. [Online]. Available: `http://arxiv.org/abs/1706.03762` (visited on 12/07/2024).

[25] D. M. Ziegler, N. Stiennon, J. Wu, *et al.*, *Fine-Tuning Language Models from Human Preferences*, arXiv:1909.08593 [cs], Jan. 2020. DOI: `10.48550/arXiv.1909.08593`. [Online]. Available: `http://arxiv.org/abs/1909.08593` (visited on 12/10/2024).

[26] Y. Bai, A. Jones, K. Ndousse, *et al.*, *Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback*, arXiv:2204.05862 [cs], Apr. 2022. DOI: `10.48550/arXiv.2204.05862`. [Online]. Available: `http://arxiv.org/abs/2204.05862` (visited on 03/26/2025).

[27] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, "Deep Reinforcement Learning from Human Preferences," in *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017. [Online]. Available: `https://papers.nips.cc/paper_files/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html` (visited on 12/10/2024).

[28] D. Ganguli, L. Lovitt, J. Kernion, *et al.*, *Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned*, arXiv:2209.07858 [cs], Nov. 2022. DOI: `10.48550/arXiv.2209.07858`. [Online]. Available: `http://arxiv.org/abs/2209.07858` (visited on 12/12/2024).

[29] L. Ouyang, J. Wu, X. Jiang, *et al.*, *Training language models to follow instructions with human feedback*, arXiv:2203.02155 [cs], Mar. 2022. DOI: `10.48550/arXiv.2203.02155`. [Online]. Available: `http://arxiv.org/abs/2203.02155` (visited on 12/11/2024).

[30] N. Stiennon, L. Ouyang, J. Wu, *et al.*, *Learning to summarize from human feedback*, arXiv:2009.01325 [cs], Feb. 2022. DOI: `10.48550/arXiv.2009.01325`. [Online]. Available: `http://arxiv.org/abs/2009.01325` (visited on 03/26/2025).

[31] Rafael Rafailov, Yaswanth Chittepu, Ryan Park, *et al.*, "Scaling Laws for Reward Model Overoptimization in Direct Alignment Algorithms," *ArXiv*, vol. abs/2406.02900, 2024. DOI: `10.48550/arXiv.2406.02900`.

[32] H. Dong, W. Xiong, D. Goyal, *et al.*, *RAFT: Reward rAnked FineTuning for Generative Foundation Model Alignment*, arXiv:2304.06767 [cs], Dec. 2023. DOI: `10.48550/arXiv.2304.06767`. [Online]. Available: `http://arxiv.org/abs/2304.06767` (visited on 03/30/2025).

[33] Z. Yuan, H. Yuan, C. Tan, W. Wang, S. Huang, and F. Huang, *RRHF: Rank Responses to Align Language Models with Human Feedback without tears*, arXiv:2304.05302 [cs], Oct. 2023. DOI: `10.48550/arXiv.2304.05302`. [Online]. Available: `http://arxiv.org/abs/2304.05302` (visited on 03/30/2025).

[34] Y. Bai, S. Kadavath, S. Kundu, *et al.*, *Constitutional AI: Harmlessness from AI Feedback*, arXiv:2212.08073, Dec. 2022. DOI: `10.48550/arXiv.2212.08073`. [Online]. Available: `http://arxiv.org/abs/2212.08073` (visited on 11/05/2024).

[35] OpenAI, *OpenAI o1 System Card*, en-US, 2024. [Online]. Available: `https://openai.com/index/openai-o1-system-card/` (visited on 09/24/2024).

[36] M. Mitchell, S. Wu, A. Zaldivar, *et al.*, "Model Cards for Model Reporting," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, arXiv:1810.03993 [cs], Jan. 2019, pp. 220–229. DOI: `10.1145/3287560.3287596`. [Online]. Available: `http://arxiv.org/abs/1810.03993` (visited on 12/12/2024).

[37] E. Ozoani, M. Gerchick, and M. Mitchell, *Model Card Guidebook*. [Online]. Available: `https://huggingface.co/docs/hub/model-card-guidebook` (visited on 03/30/2025).

[38] M. K. Richter, "Revealed Preference Theory," *Econometrica*, vol. 34, no. 3, pp. 635–645, 1966, Publisher: [Wiley, Econometric Society], ISSN: 0012-9682. DOI: 10.2307/1909773. [Online]. Available: https://www.jstor.org/stable/1909773 (visited on 03/28/2025).

[39] C. P. Chambers and F. Echenique, *Revealed Preference Theory* (Econometric Society Monographs). Cambridge: Cambridge University Press, 2016, ISBN: 978-1-107-08780-4. DOI: 10.1017/CBO9781316104293. [Online]. Available: https://www.cambridge.org/core/books/revealed-preference-theory/7637115A86AFE512E920B6111780FEB5 (visited on 12/11/2024).

[40] J. Barrett, "Social Reform in a Complex World," en, *Journal of Ethics and Social Philosophy*, vol. 17, no. 2, 2020, ISSN: 1559-3061. DOI: 10.26556/c47s8023. [Online]. Available: https://mail.jesp.org/index.php/jesp/article/view/900 (visited on 08/19/2024).

[41] D. Wiens, "Assessing ideal theories: Lessons from the theory of second best," en, *Politics, Philosophy & Economics*, vol. 15, no. 2, pp. 132–149, May 2016, ISSN: 1470-594X, 1741-3060. DOI: 10.1177/1470594X15620343. [Online]. Available: http://journals.sagepub.com/doi/10.1177/1470594X15620343 (visited on 08/19/2024).

[42] D. Estlund, "Utopophobia," en, *Philosophy & Public Affairs*, vol. 42, no. 2, pp. 113–134, Mar. 2014, ISSN: 0048-3915, 1088-4963. DOI: 10.1111/papa.12031. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/papa.12031 (visited on 09/23/2024).

[43] R. E. Goodin, "Political Ideals and Political Practice," en, *British Journal of Political Science*, vol. 25, no. 1, pp. 37–56, Jan. 1995, ISSN: 0007-1234, 1469-2112. DOI: 10.1017/S0007123400007055. [Online]. Available: https://www.cambridge.org/core/product/identifier/S0007123400007055/type/journal_article (visited on 08/19/2024).

[44] R. G. Lipsey and K. Lancaster, "The General Theory of Second Best," *The Review of Economic Studies*, vol. 24, no. 1, pp. 11–32, 1956, Publisher: [Oxford University Press, Review of Economic Studies, Ltd.], ISSN: 0034-6527. DOI: 10.2307/2296233. [Online]. Available: https://www.jstor.org/stable/2296233 (visited on 12/12/2024).

[45] Z. Lipton, J. McAuley, and A. Chouldechova, "Does mitigating ML' s impact disparity require treatment disparity?" In *Advances in Neural Information Processing Systems*, vol. 31, Curran Associates, Inc., 2018. [Online]. Available: https://papers.nips.cc/paper_files/paper/2018/hash/8e0384779e58ce2af40eb365b318cc32-Abstract.html (visited on 12/12/2024).

[46] C. Castro, "What's Wrong with Machine Bias," en, *Ergo, an Open Access Journal of Philosophy*, vol. 6, no. 20201214, Jul. 2019, ISSN: 2330-4014. DOI: 10.3998/ergo.12405314.0006.015. [Online]. Available: http://hdl.handle.net/2027/spo.12405314.0006.015 (visited on 01/18/2024).

[47] S. Corbett-Davies, J. D. Gaebler, H. Nilforoshan, R. Shroff, and S. Goel, *The Measure and Mismeasure of Fairness*, arXiv:1808.00023 [cs], Aug. 2023. DOI: 10.48550/arXiv.1808.00023. [Online]. Available: http://arxiv.org/abs/1808.00023 (visited on 12/12/2024).

[48] Anka Reuel and Devin Ma, "Fairness in Reinforcement Learning: A Survey," *ArXiv*, vol. abs/2405.06909, 2024. DOI: 10.48550/arXiv.2405.06909.

[49] J. Kleinberg, S. Mullainathan, and M. Raghavan, *Inherent Trade-Offs in the Fair Determination of Risk Scores*, arXiv:1609.05807 [cs], Nov. 2016. DOI: 10.48550/arXiv.1609.05807. [Online]. Available: http://arxiv.org/abs/1609.05807 (visited on 12/10/2024).

[50] I. Gabriel, "Artificial Intelligence, Values, and Alignment," en, *Minds and Machines*, vol. 30, no. 3, pp. 411–437, Sep. 2020, ISSN: 0924-6495, 1572-8641. DOI: 10.1007/s11023-020-09539-2. [Online]. Available: https://link.springer.com/10.1007/s11023-020-09539-2 (visited on 10/29/2024).

[51] S. Casper, X. Davies, C. Shi, *et al.*, *Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback*, en, arXiv:2307.15217 [cs], Sep. 2023. [Online]. Available: http://arxiv.org/abs/2307.15217 (visited on 08/11/2024).

[52] Shreyas Chaudhari, Pranjal Aggarwal, Vishvak Murahari, *et al.*, "RLHF Deciphered: A Critical Analysis of Reinforcement Learning from Human Feedback for LLMs," *ArXiv*, vol. abs/2404.08555, 2024. DOI: 10.48550/arXiv.2404.08555.

[53] Z. Sun, Y. Shen, Q. Zhou, *et al.*, *Principle-Driven Self-Alignment of Language Models from Scratch with Minimal Human Supervision*, arXiv:2305.03047 [cs], Dec. 2023. DOI: 10.48550/arXiv.2305.03047. [Online]. Available: http://arxiv.org/abs/2305.03047 (visited on 03/30/2025).

[54] M. Y. Guan, M. Joglekar, E. Wallace, *et al.*, *Deliberative Alignment: Reasoning Enables Safer Language Models*, arXiv:2412.16339 [cs], Jan. 2025. DOI: 10.48550/arXiv.2412.16339. [Online]. Available: http://arxiv.org/abs/2412.16339 (visited on 02/03/2025).

[55] S. Schroeder, *Microsoft 'deeply sorry' for offensive tweets by its Tay chatbot*, en, Section: Tech, 2016. [Online]. Available: https://mashable.com/article/tay-offensive-tweets (visited on 12/11/2024).

[56] D. Lee, *Tay: Microsoft issues apology over racist chatbot fiasco - BBC News*, 2016. [Online]. Available: https://www.bbc.com/news/technology-35902104 (visited on 12/11/2024).

[57] E. Anderson, *The Imperative of Integration.* Princeton University Press, 2010, ISBN: 978-0-691-13981-4. [Online]. Available: http://www.jstor.org/stable/j.ctt7t225 (visited on 12/11/2024).

[58] Utkarsh Agarwal, K. Tanmay, Aditi Khandelwal, and Monojit Choudhury, "Ethical Reasoning and Moral Value Alignment of LLMs Depend on the Language We Prompt Them in," pp. 6330–6340, 2024, Num Pages: 6330-6340. DOI: 10.48550/arXiv.2404.18460.

[59] Vamshi Bonagiri, Sreeram Vennam, Manas Gaur, and P. Kumaraguru, "Measuring Moral Inconsistencies in Large Language Models," *ArXiv*, vol. abs/2402.01719, 2024. DOI: 10.48550/arXiv.2402.01719.

[60] Vamshi Bonagiri, Sreeram Vennam, Priyanshul Govil, P. Kumaraguru, and Manas Gaur, "SaGE: Evaluating Moral Consistency in Large Language Models," *ArXiv*, vol. abs/2402.13709, 2024. DOI: 10.48550/arXiv.2402.13709.

[61] A. J. Simmons, "Ideal and Nonideal Theory," en, *Philosophy & Public Affairs*, vol. 38, no. 1, pp. 5–36, Jan. 2010, ISSN: 00483915, 10884963. DOI: 10.1111/j.1088-4963.2009.01172.x. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/j.1088-4963.2009.01172.x (visited on 08/19/2024).

[62] L. Valentini, "A Paradigm Shift in Theorizing About Jusitce? A critique of Sen," en, *Economics and Philosophy*, vol. 27, no. 3, pp. 297–315, Nov. 2011, ISSN: 0266-2671, 1474-0028. DOI: 10.1017/S0266267111000228. [Online]. Available: https://www.cambridge.org/core/product/identifier/S0266267111000228/type/journal_article (visited on 08/19/2024).

[63] G. F. Gaus, *The tyranny of the ideal: justice in a diverse society*, eng. Princeton: Princeton university press, 2016, ISBN: 978-0-691-15880-8.

[64] L. Jiang, J. D. Hwang, C. Bhagavatula, *et al.*, *Can Machines Learn Morality? The Delphi Experiment*, arXiv:2110.07574 [cs], Jul. 2022. DOI: 10.48550/arXiv.2110.07574. [Online]. Available: http://arxiv.org/abs/2110.07574 (visited on 12/06/2024).

[65] N. Bostrom, *Superintelligence: paths, dangers, strategies*, eng, Reprinted with corrections. Oxford: Oxford University Press, 2017, ISBN: 978-0-19-873983-8.

[66] R. Kurzweil, *The singularity is nearer: when we merge with Al.* New York: Viking, 2024, OCLC: on1438926317, ISBN: 978-0-399-56276-1.

[67] F. Gursoy and I. A. Kakadiaris, *System Cards for AI-Based Decision-Making for Public Policy*, arXiv:2203.04754 [cs], Aug. 2022. DOI: 10.48550/arXiv.2203.04754. [Online]. Available: http://arxiv.org/abs/2203.04754 (visited on 12/12/2024).

[68] W.-L. Chiang, L. Zheng, Y. Sheng, *et al.*, *Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference*, arXiv:2403.04132 [cs], Mar. 2024. DOI: 10.48550/arXiv.2403.04132. [Online]. Available: http://arxiv.org/abs/2403.04132 (visited on 12/12/2024).

[69] E. Tennant, S. Hailes, and M. Musolesi, *Moral Alignment for LLM Agents*, en, arXiv:2410.01639 [cs], Oct. 2024. [Online]. Available: http://arxiv.org/abs/2410.01639 (visited on 10/20/2024).

[70] A. Askell, Y. Bai, A. Chen, *et al.*, *A General Language Assistant as a Laboratory for Alignment*, arXiv:2112.00861 [cs], Dec. 2021. DOI: 10.48550/arXiv.2112.00861. [Online]. Available: http://arxiv.org/abs/2112.00861 (visited on 03/30/2025).

[71] *Claude's Constitution*, en. [Online]. Available: https://www.anthropic.com/news/claudes-constitution (visited on 11/05/2024).

[72] *Collective Constitutional AI: Aligning a Language Model with Public Input*, en. [Online]. Available: https://www.anthropic.com/news/collective-constitutional-ai-aligning-a-language-model-with-public-input (visited on 11/05/2024).

[73] I. Gabriel, "Toward a Theory of Justice for Artificial Intelligence," en, *Daedalus*, vol. 151, no. 2, pp. 218–231, May 2022, ISSN: 0011-5266, 1548-6192. DOI: 10.1162/daed_a_01911. [Online]. Available: https://direct.mit.edu/daed/article/151/2/218/110610/Toward-a-Theory-of-Justice-for-Artificial (visited on 03/27/2025).

[74] M. Adams, "An ideology critique of nonideal methodology," en, *European Journal of Political Theory*, vol. 20, no. 4, pp. 675–697, Oct. 2021, ISSN: 1474-8851, 1741-2730. DOI: 10.1177/1474885119858727. [Online]. Available: http://journals.sagepub.com/doi/10.1177/1474885119858727 (visited on 09/23/2024).

[75] A. Berg, "Incomplete Ideal Theory," en, *Social Theory and Practice*, vol. 45, no. 4, pp. 501–524, 2019, ISSN: 0037-802X. DOI: 10.5840/soctheorpract2019122069. [Online]. Available: http://www.pdcnet.org/oom/service?url_ver=Z39.88-2004&rft_val_fmt=&rft.imuse_id=soctheorpract_2019_0045_0004_0501_0524&svc_id=info:www.pdcnet.org/collection (visited on 08/19/2024).

[76] I. Robeyns, "Are transcendental theories of justice redundant?" en, *Journal of Economic Methodology*, vol. 19, no. 2, pp. 159–163, Jun. 2012, ISSN: 1350-178X, 1469-9427. DOI: 10.1080/1350178X.2012.683587. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/1350178X.2012.683587 (visited on 08/19/2024).

[77] T. Shelby, *Dark ghettos: injustice, dissent, and reform*, eng. Cambridge, Massachusetts London, England: The Belknap Press of Harvard University Press, 2016, ISBN: 978-0-674-98407-3 978-0-674-97050-2.