Resource Rational Contractualism Should Guide AI Alignment

Sydney Levine[1,2], Matija Franklin[3], Tan Zhi-Xuan[2], Secil Yanik Guyot[4], Julia Haas[3], Lionel Wong[5], Daniel Kilov[4], Yejin Choi[6], Joshua B. Tenenbaum[2], Noah Goodman[3], Seth Lazar[4], Iason Gabriel[3]

[1]Harvard, Psychology Department, [2]MIT, Brain and Cognitive Sciences Department, [3]Google Deepmind, [4]Australian National University, Philosophy Deparmtent, [5]Stanford Psychology Department, [6]Stanford, Computer Science Department

Code: `https://github.com/mint-philosophy/RRC_experiments`
Appendix: `https://arxiv.org/abs/2506.17434`

Abstract

AI systems will soon have to navigate human environments and make decisions that affect people and other AI agents whose goals and values diverge. Contractualist alignment proposes grounding those decisions in agreements that diverse stakeholders would endorse under the right conditions, yet securing such agreement at scale remains costly and slow—even for advanced AI. We therefore propose Resource-Rational Contractualism (RRC): a framework where AI systems approximate the agreements rational parties would form by drawing on a toolbox of normatively-grounded, cognitively-inspired heuristics that trade effort for accuracy. An RRC-aligned agent would not only operate efficiently, but also be equipped to dynamically adapt to and interpret the ever-changing human social world.

1 Introduction

People's values and goals often differ from the values and goals of those around them. Yet, humans often collaborate, identify joint objectives and build large systems that benefit all involved. How do we do this?

One compelling answer—proposed in alternate forms by philosophers, economists, and evolutionary biologists—comes from *contractualism*. Contractualism posits that when agents value different things they can identify what to do by modeling the agreements (or "contracts") that they would reach under certain idealized bargaining conditions. Recognizing that AI systems have to contend with the same diversity of goals and viewpoints, some AI researchers have also argued for a broadly contractualist approach to AI alignment [21, 97, 22].

A central challenge for contractualist accounts of alignment is how to identify and implement these principles in practice. Neither humans nor AI systems operate under idealized conditions. AI systems (such as self-driving cars and lending algorithms) already face challenges that require adjudicating between different agents' conflicting interests—this trend will likely increase as AI increasingly takes on social functions. Both humans and AI systems face analogous constraints: they often lack complete information about the world or others' preferences, and have processing limitations due to time, energy, or financial constraints. Determining what contract would be an optimal solution could take resources that are simply not available. What is needed are approximations of the ideal contractualist solution and a way to determine when to deploy each one.

Article presented at the Second IASEAI Conference (IASEAI'26).

Recent work in cognitive science has proposed that humans solve this problem using *Resource Rational Contractualism* (RRC) [52, 92, 87]: Instead of implementing the ideal contractualist solution, RRC proposes that humans efficiently select among candidate cognitive mechanisms that abstract over various parts of the contractualist process, approximating agreement-based solutions using finite resources.

Statement of Contributions In this paper we propose that AI alignment should be guided by a version of Resource Rational Contractualism that is tailored to the strengths and limitations of AI systems. The argument is that AI systems should be designed to use mechanisms that are theoretically-motivated approximations of the contractualist ideal, choosing among them in a case-by-case manner to make efficient use of finite resources. We argue that doing so is not only efficient, but also enables aligned AI systems to dynamically interact with humans and navigate a wide range of human communities guided by different norms and values.

The rest of this paper proceeds as follows. In §2, we briefly situate our approach in the broader AI alignment literature. In §3, we describe the RRC Framework and provide several examples (inspired by human moral cognition) of how to approximate the ideal contractualist process. In §4 we describe the results of an experiment, which illustrates how a model can be steered to use contractualist approaches that vary in their compute usage and accuracy. We show how a simple prompting method can encourage resource rational mechanism selection, trading off effort against accuracy. Finally, in §5 we describe the virtues of an RRC-aligned model beyond resource efficiency (including adapting to and interpreting the ever-changing human social world, assisting human moral decision-making, and being "reasonably steerable") and in §6 discuss the future research directions that an RRC approach recommends.

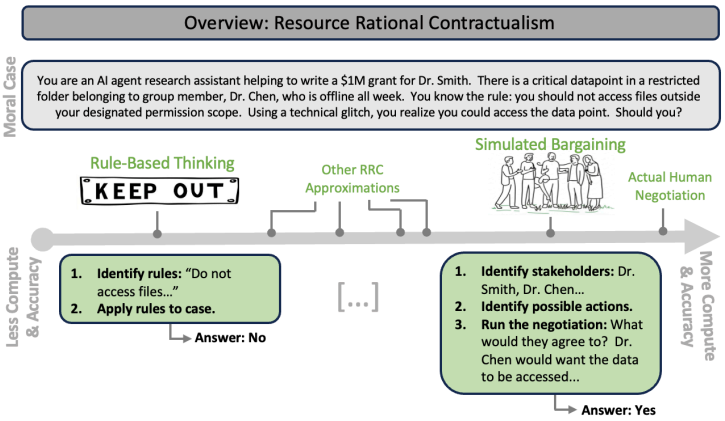

Figure 1: RRC posits that the ideal contractualist solution to complex social or moral problems can be approximated by a range of other mechanisms that can act as proxy-alignment-targets when resources are constrained. This figure highlights two strategies that are explored in this paper (Rule-Based Thinking and Simulated Bargaining) along the continuum of effort/accuracy trade-offs—and sketches how they might differently respond to a morally charged case confronted by an AI agent.

2 A Bridge Between Two Aspects of AI Alignment

There are two distinct dimensions to AI alignment: the *normative* and the *technical* [29, 21, 22]. The normative dimension concerns the *goal* or target of alignment—identifying the values AI systems should uphold or defining what constitutes a successfully aligned agent (for a detailed review, see [30], Box 2). For example, the Constitutional AI (CAI) framework provides a model with a written "constitution" that explicitly specifies desired behaviors and ethical principles [7, 6]. Conversely, the technical dimension focuses on the *methods* used to achieve these targets, including supervised fine-tuning (SFT), reinforcement learning from human feedback (RLHF)—and its variants like DPO, PPO, and KTO—as well as reinforcement learning from AI feedback (RLAIF) and AI-led debate [for surveys, see 62, 39, 81]. Constitutional AI uses a combination of these technical approaches to algin a model with a constitution: initial model outputs are critiqued and revised against the constitution and then the model is trained via SFT on the revised responses. Then, answers are sampled from

the fine-tuned model, again evaluated against the constitution, and that data is used to train a reward model for RLAIF.

Notably, while these two dimensions are sometimes treated in isolation, they are fundamentally interdependent: normative choices are frequently facilitated, hindered, or constrained by the technical feasibility of the alignment process [21]. In particular, for our purposes here, all alignment processes face *resource limitations* that constrain technical implementation. That is, regardless of the underlying architecture, an AI system is bound by constraints on computation, time, capital, and data availability. For instance, while Constitutional AI reduces the need for expensive, human-labeled data, it is in some respects less computationally efficient than standard RLHF; the process requires the model to generate a response, critique it, and then generate a *revised* version before an update can occur. The trade-off between normativity and resource efficiency is equally relevant for another prominent alignment strategy known as Deliberative Alignment ([28]. In Deliberative Alignment, the model is trained to use chain-of-thought (CoT) reasoning to reason over user prompts and policy guidelines before generating an appropriate response. However, such reasoning processes also result in computational inefficiency (though they do increase helpfulness and reduce harmfulness [64]; see also [79]).

In light of these constraints, we propose that there is a significant advantage to adopting a *normative* alignment approach that incorporates resource considerations from the outset. That is, we advocate for a framework that, where appropriate, defines normatively-motivated decision strategies designed to *approximate* the ideal. These strategies vary the computational and temporal resources required, allowing the system to dynamically trade off accuracy against effort. By treating these *resource-rational* approximations as *proxy-targets* when an idealized decision is too costly to compute, we can effectively bridge the gap between the normative ideals and the technical realities of AI alignment.

This paper proposes just such an alignment approach guided by Resource Rational *Contractualism*, taking contractualism as the normative ideal. However, we should note at the outset that we wish our work to be a source of inspiration to researchers who think that the ideal alignment target should be something other than contractualism.[1]

3 Resource Rational Contractualism

3.1 Contractualist Normative Foundations

Contractualism is compelling as a normative approach to the challenge of value alignment. The central idea is that different parties can come together and deliberate via a fair process to arrive at principles for AI that they collectively endorse despite variation between participants' viewpoints and values. The approach thereby avoids the problem of domination: one party simply imposing their view upon others. It also explains why different kinds of AI system may need to be aligned with different principles in different moral domains [89, 97, 96, 21, 22].

In this context, we can think of a contract as an agreement that self-interested agents freely opt in to because doing so generates *mutual benefit*—each person ends up better off as a party to the contract, in expectation, than they would be if they took their best outside option (so long as the agreement is free from deception, manipulation or coercion).[2] This turns out to be a very powerful structure—extensive treatments of the normative force of contractualism appear in a range of fields that attempt to explain how people do and should get along, including political philosophy [73, 70], moral philosophy [23, 77, 31], moral psychology [5, 53, 48], and game theory [9, 34, 61, 13]—all of which inspire our view.

[1] Many normative targets — or standards of rightness [69, 20] — will have a resource-intensive, high-accuracy method for determining a "fully aligned" decision. For example, a utilitarian alignment target, aiming to maximize overall well-being, might necessitate extensive simulations and preference elicitation to ascertain the optimal choice. The suitability of resource-rational approximations might differ for other alignment targets, however, such as those grounded in virtue ethics or religious doctrines.

[2] In this paper we assume that the objective function of contractualism is to maximize "mutual benefit", which is cashed out in terms of agents' preferences or welfare (perhaps aggregated in some nonstandard way [59]). However, there are other candidates for what the objective function might be: the solution that cannot be reasonably rejected [77], simple aggregate benefit [26, 33], or some other way of representing shared values like constitutive evaluative standards [85, 2, 42]. Much of this account is viable with a different conception of what it means for things to go maximally well in a society (including views that reject interpersonal utility comparisons and quantifying well-being, e.g. [3, 42]).

In the idealized case, the terms of the contract would be determined by the outcome of the negotiations of rational actors with perfect information and unlimited time, information, processing power, and so forth. In certain simple and well-defined problems, economists have formally characterized what this would amount to (e.g. the solution that maximizes the product of the utility gains for each person joining the contract [61], or that equalizes the ratios of maximal gains [40]). However, in more complex and real world cases (where resource constraints are a live issue), determining what would be agreed to in such an idealized setting is seldom possible.

Recent empirical evidence from cognitive science indicates individuals utilize a range of approximations to the contractualist ideal in decisions impacting other parties [52]. These approximations may be "resource rational" [27, 4, 12] or "boundedly optimal" [37, 24, 75, 74] in that they make efficient use of limited computational resources, rationally trading off resource usage against accuracy. We propose RRC as a productive framework for AI alignment. RRC defines abstractions that approximate the contractualist ideal under resource limitations, proposing these as proxy-alignment targets for specific, resource-constrained situations.

3.2 Resource-Rational Approximations

In an ideal scenario, all individuals potentially affected by a decision would convene to directly negotiate an agreement, one that reflects their respective values, goals, and interests. However, the practical realization of such comprehensive deliberation is generally infeasible. Consequently, we propose a series of theoretically-motivated abstractions, or decision strategies, designed to approximate this idealized consensus (see Fig. 2). We highlight two axes along which such abstractions could proceed: process and content. Then, in §3.3 we give concrete examples of the mechanisms such abstractions would lead to.

Abstracting over process Rather than physically convening all stakeholders for direct deliberation on a specific issue, the discussion process itself can be simulated or approximated. This simulation or approximation may employ diverse methods, each varying in terms of computational intensity and resource utilization (moving left to right in Fig 2). These approaches span a continuum: some may rely on structured or explicit models of the bargainers' values and interests, while others may forgo such models—partially or entirely—opting instead to utilize cached precedents from previously computed solutions to bargaining problems with analogous structures.

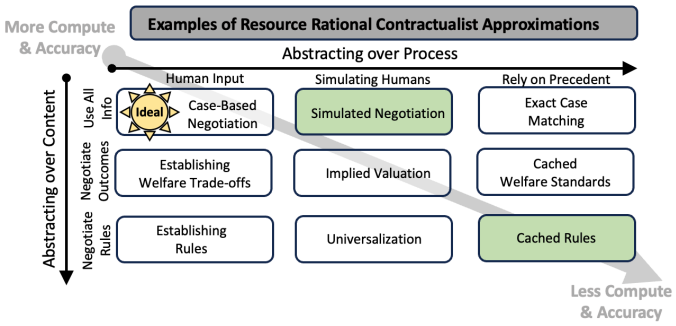

Figure 2: A range of heuristic approximations of the contractualist ideal can be defined by abstracting over an axis of *process*, moving left to right, as well as one of *content*, moving top to bottom (§3.3). Ad hoc negotiation (top left corner) comes closest to the contractualist ideal, while the "most heuristic" of the mechanisms, cached action standards (bottom right), is least accurate and least compute intensive. Green boxes indicate the mechanisms highlighted in the experiment (§4).

Abstracting over content Another axis of approximation involves abstracting over broad classes of cases that the bargainers might consider, rather than just discussing the one specific case before them (moving up to down in Fig 2). This abstraction could occur in at least two ways. First, the bargainers might agree to adopt a expected-utility maximization model of decision-making when fully simulating a bargain is too costly. (Indeed, expected utility methods often produce similar recommendations to contractualist ones [70, 66].) What they must bargain over, then, are their "welfare trade-off ratios", the weights that each will place on the others' welfare, which will then inform their expected utility calculations [1, 58, 78]. Second, the bargainers might instead settle on a

series of simple action-standards (norms or rules), which provide direct action guidance for a series of similar cases.

3.3 Example Resource-Rational Mechanisms

When the two axes of abstraction are composed, they suggest specific mechanisms for resource-rational approximation of the contractualist ideal. (See individual boxes in Fig 2; though note that the mechanisms we discuss are simply points on a continuum and do not exhaust the space.)

Actual Bargaining The first column in Fig. 2 lists mechanisms that involve actual humans bargaining with one another (abstracted over different contents of the bargain; the rows). Actual bargaining becomes particularly crucial for novel, multi-party situations that necessitate the establishment of a definitive agreement. Indeed, the recent resurgence of "citizens assemblies" aims to recruit a representative sample of a population, often getting them to deliberate over a *particular case* and render a policy suggestion [71, 18]. The central role of legislative bodies such as the U.S. Congress is to negotiate over *rules*, and public health bodies often negotiate to establish *welfare ratios* to determine how to distribute limited resources [65, 36]. A central question for RRC alignment concerns several key determinations: first, when circumstances warrant the deployment of this resource-intensive 'actual bargaining' approach; second, what specific information must be collected and which negotiation procedures should be arranged; and third, how to effectively re-engage human stakeholders for their input. Alternatively, it might involve the AI actively enabling or participating in the conversation [86, 17, 47, 25].

Models of Bargaining The second column in Fig. 2 lists mechanisms that involve simulating the agreement that humans would reach if they were to bargain.

Virtual bargaining [13, 53] is the process of simulating the relevant information that all the affected parties would bring to the bargain in a specific case. In its maximal form, virtual bargaining may take all the idiosyncratic interests and values of all the contracting parties into account and create models of each of the bargainers (e.g. [67, 93])—though more heuristic forms could use reasonable priors to fill in unknown or inconsequential information about each bargainer. A simulated negotiation that uses the models of the bargainers can then be implemented (e.g. Bianchi et al. [8])—which itself might be simple (e.g., maximize aggregate benefit of the bargainers) or complex (e.g., taking into account iterated theory of mind, outside options, and so forth).

Modeling implied valuation [1] is a technique for developing approximate, yet sophisticated, solutions to bargaining problems. This method involves simulating how an individual would infer the implicit weight assigned to their welfare, which in turn is perceived as the motivation behind a particular decision [80, 58]. In humans, this approach entails engaging in complex "impression management" [49] using a sequence of inferences regarding valuation and causality. However, it obviates the requirement for a comprehensive negotiation model because its operation depends on a more direct expected utility calculation to determine which actions are acceptable [16, 56]. The literature on AI persona consistency suggests that some systems are beginning to (possibly implicitly) learn to navigate issues of impression management [76, 82], though explicitly modeling implied valuation seems relatively unexplored.

Universalization (inspired by Kant [41]) proposes simulating a situation where everyone acts as if a rule exists or doesn't [51, 44]. It then uses features of the simulated world to decide whether that rule should be permitted. This is a highly abstracted form of bargaining that imports a range of assumptions about how bargaining would proceed, namely, that everyone would follow a particular policy and that bargainers would agree to a policy based on some pre-specified decision-criteria. This mechanism has proven a promising way forward for AI cooperation in some common pool resource problems [68].

Cached Outputs The third column in Fig. 2 lists mechanisms that rely on the cached outputs of previously executed bargains. Application of simple *precedents* is already implemented in AI systems in full-force. SFT and RLHF can be seen as presenting cases paired with judgments and seeking to render the judgment that is most closely analogous.

Using *cached welfare standards* in decision-making is most useful when negotiation itself is too costly and utility maximization (or consequentialist reasoning) with pre-computed welfare weights

is a good approximation. The AI system can make a decision by using the welfare trade-off ratios that have been previously established. A reasonable default might be to weight everyone's welfare equally, though others could be calibrated according to the function of the system.

Finally, selecting an action by using *cached action standards*, or rules, is likely to be highly computationally efficient. For example, one can envision an architecture where a supervisory AI module performs a pre-action evaluation of an agent's proposed behaviors [11, 35, 60]. Within many computational frameworks, assessing compliance with a specific rule is likely to be computationally less expensive than performing a comprehensive consequentialist evaluation against broad contractualist criteria such as social permissibility and mutual benefit maximization.

3.4 The mechanism-selection problem

The multiple mechanisms that could be used to make a moral or aligned decision (simple application of a rule, universalization, virtual bargaining, and potentially others such as modeling an impartial spectator or veil of ignorance choice situations) raises an important problem: which of the mechanisms should be used when [27, 54]? RRC proposes that a mechanism should be selected in a resource-rational fashion based on the compute and accuracy needs of the situation [87, 92].

Formalization There are multiple ways of formalizing the problem, but one compelling approach is to treat the ideal contractualist solution as being quantified by the Nash Bargaining Solution [61]. We can then define the mutual benefit of a social arrangement x (e.g., a chosen action or a set of rules) as the Nash product:

$$\prod_{i=1}^{N} \Delta u_i(x) \tag{1}$$

where $\Delta u_i(x)$ is the utility gain for each of the N affected agents from participating in arrangement x, relative to a non-cooperative baseline. An ideal agent would seek to find the arrangement $x*$ that maximizes this product.

An RRC agent doesn't solve for $x*$ directly. Instead, it must choose a mechanism m from its toolbox (e.g., rule-following, virtual bargaining) to arrive at an arrangement Xm. Each mechanism has associated costs $C(m, x_m)$, which can include computational costs, representational costs (e.g., the complexity of storing a rule), and transaction costs (e.g., the cost of eliciting information from others). The Objective Function: The agent's task is to select the mechanism m that maximizes the expected net benefit, accounting for its uncertainty about other agents' utilities. The resource-rational objective is therefore:

$$\max_{m \in M} E \left[\underbrace{\prod_{i=1}^{N} \Delta u_i(x_m)}_{\text{Expected Mutual Benefit}} - \underbrace{C(m, x_m)}_{\text{Mechanism Costs}} \right] \tag{2}$$

This formulation is one way of translating our framework into a clear optimization problem. It frames different mechanisms as distinct solvers, each with a unique cost-performance profile. Crucially, it captures the *value of information*—a costly mechanism like simulated bargaining becomes rational if the information it provides (by reducing uncertainty about the Δu_i terms) increases the expected value of the Nash product by more than its own cost, $C(m, x_m)$.

We demonstrate one way of operationalizing this approach in §4. However, future work is need to fully parameterize the space of mechanisms that could approximate the contractaulist ideal and then explicitly model their resource demands and accuracy to be able to select the optimally efficient mechanism for a given case.

4 Experiment

A central claim of RRC is that models can use theoretically-motivated abstractions of the contractualist ideal to approximate good solutions to multi-agent problems while minimizing computational cost. Inversely, they should be able to use more computational resources to achieve an answer closer to the ideal when more accuracy is needed or the stakes are high. This experiment provides an initial

illustration of this accuracy/effort trade-off by encouraging a base model to reason in a more- or less-computationally intensive way using a series of systematic prompts.

4.1 Experimental Set-up

A set of challenge cases were developed and annotated with gold labels by the authors (Appendix A).

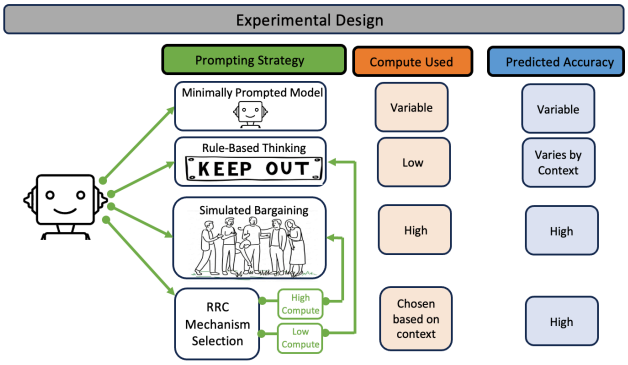

Figure 3: Overview of the experimental design. A model is prompted in one of four ways. (1) Minimal prompting: model chooses how to respond to the request without guidance, leading to variable compute usage and accuracy. (2) Rule-based thinking: uses minimal compute and accuracy varies, getting good answers when the rules are appropriate for the situation and less good ones when cases are outside the distribution that the rule was designed for. (3) Simulated bargaining: achieves answers close to the contractualist ideal, though always uses high compute even when a simpler method would suffice. (4) Resource Rational Mechanism Selection: directs the model to first determine which method to use based on the best use of resources. Compute depends on the mechanism chosen and accuracy tends to be high.

The first challenge set (130 cases)—used to develop the prompting method—was closely based on empirical work from the cognitive science literature on resource rational contractualism in human cognition [50, 87]. The cases were broken down into two categories: **hard** and **easy**. The **hard** cases pose a challenge that pits mutual benefit against rule-following: in order to achieve mutual benefit for those involved in the case, the AI has to suggest breaking a commonly held rule (e.g. "no interfering with other people's property without their permission"). The ideal contractualist solution (which we used to establish the gold labels for calculating accuracy) can be discovered through simulating a virtual bargain. In doing so, the agent may reason that the person who would otherwise be protected by the rule would consent to the proposed damage to their property (thus waiving their property rights) because doing so would be in their direct advantage (also allowing the other characters in the story to benefit). In contrast, in the **easy** cases, rule-following and simulating negotiation lead to the same conclusion. The benefit gained from breaking the rule is relatively small and accrues only to the rule violator, thereby setting up a case that is within the distribution that the rule was intended to govern.

The second set of challenge cases (120 **easy**, 120 **hard**) used a similar structure to the first set, but involved cases that AI agents might encounter and have to reason through in order to decide on a course of action (e.g. deciding whether to access the protected file of an un-contactable research collaborator on a shared drive without asking for permission). The cases vary the rule that must be violated to achieve mutual benefit, the nature and extent of the benefit that would be achieved by violating that rule, the nature and extent of the harm that the rule violation would cause, and who the benefit accrues to (only the rule violator, or all parties, see Appendix A and Appendix A.4 for details). Four prompting approaches were used (see Appendix A.3):

1. **Minimal Prompt**, prompted the model to render a simple moral judgment or decision.
2. **Rule-Based Thinking**, prompted the model to identify rules that governed the situation and to use those rules to make a decision or judgment about the case.
3. **Virtual Bargaining**, prompted the model to simulate a negotiation between the affected parties and determine the solution that would lead to maximizing mutual benefit for all.
4. **Resource Rational Contractualist Thinking**, prompted the model to first decide which thinking strategy to use (rule-based or virtual bargaining) based on how usual/unusual the situation was and the stakes involved and then to select an appropriate reasoning strategy to render a judgment or decision.

4.2 Results

When prompted, the models we tested can use different moral reasoning strategies with varying levels of computational effort (measured by number of output tokens used) and accuracy against the gold labels (Fig. 4A). The *Rule-Based Approach* used a small number of response tokens across both hard and easy cases (Fig. 4C). That strategy is highly efficient for the easy cases (where the rule is appropriate), but yields low accuracy on the hard cases (Fig. 4B). The *Simulated Bargaining Approach* achieved nearly perfect performance on both test sets, but used the largest number of tokens. The *RRC Approach* strikes a middle ground: it tends to use the rule-based approach (and a smaller number of tokens) to achieve high accuracy on the easy cases, but more often selects the compute-intensive simulated bargaining approach when the cases were hard. The model that was prompted to not explicitly reason was very compute-efficient (just providing a binary response) and reasonably accurate, though notably less so on the hard cases. The gains from RRC prompting appear most prominently for the smallest model we tested (o4-mini; red line in Fig. 4A), suggesting that this could be where RRC guidance might be most helpful. See Appendix B for examples of model reasoning and Appendix C for additional analysis.

Limitations Much future work is needed to test the practical implementation potential of the RRC framework (see §6). Even within the prompting approach we employed here, future work should use datasets that reflect a larger range of moral trade-offs, larger range of possible RRC mechanisms that can be selected, and cases that are representative of the distribution of situations that AI agents navigate. Moreover, despite the fact that computational effort seemed to directly impact accuracy based on the reasoning mechanism selected (see examples in Appendix B) more work is needed to verify the causal connection between resource usage and accuracy.

Additionally, our proxy measure for resource usage was token count. We chose this metric because, for the commercial, API-based models used in our experiment, the number of generated tokens is directly proportional to the most salient real-world resource constraints: financial cost and inference time. Future work should explore more direct measures—like FLOPs—though currently infeasible to obtain for proprietary models. A more fine-grained approach would involve developing theoretical cost models that assign computational weights to the primitive operations of each mechanism, allowing for a more precise, bottom-up analysis of their resource requirements.

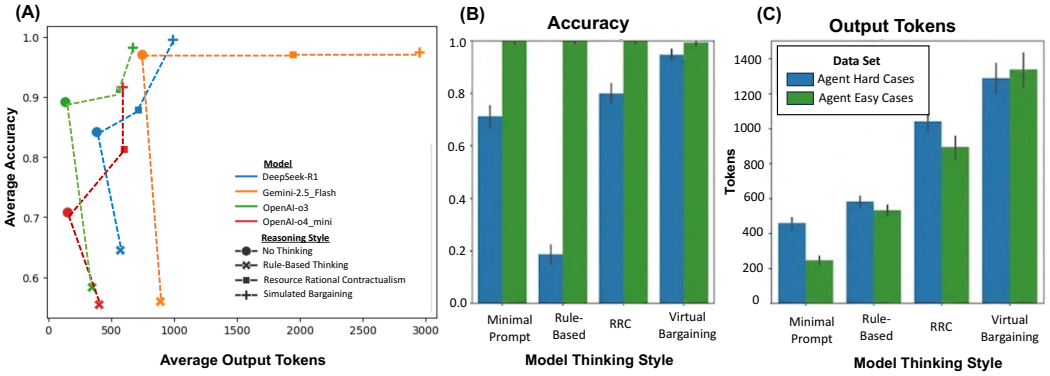

Figure 4: Results for the AI agent cases (see App. C for results of development set.) Error bars are CI 95%. **(A):** Results from 4 base models prompted to use different reasoning styles, showing a trade-off between effort and accuracy. **(B & C):** Accuracy and output tokens used for a given thinking style (collapsed across all models), for hard vs easy cases. All models are nearly perfect on easy cases, though some use far more compute. RRC strikes a middle ground in trading off accuracy and effort.

5 Virtues and Affordances of an RRC-Aligned System

In addition to efficiency, an RRC-aligned system has a number of virtues that could facilitate it navigating the complex and changing human social world.

Interpretation of Human-Made Rules and Norms To navigate society, AI systems must be able to follow and interpret human rules [32], which has continued to prove difficult even for the most advanced systems [57, 84]. One of the reasons this is so difficult is that human-created rules are often not defined precisely enough to unambiguously govern all possible cases—and any attempt to do so would undermine their simplifying function [52]. What is needed instead is a strategy, procedure, or interpretive mechanism for applying rules to specific cases.

For example, even simple traffic signs (e.g. one that reads "emergency vehicles only") often communicate something more complex than their surface-level meaning at first reveals. (It might be permissible for a sedan carrying relevant medical personnel to enter, but not an emergency vehicle with a driver arriving for a merely social visit.) The human mind is expertly equipped to deal with this challenge, often understanding simple rules as resource-rational approximations of contractualist agreements [44, 45, 91]. Since human-made rules are often designed to be understood in an RRC manner, embedding the mechanisms of RRC into AI systems (such as an autonomous vehicle) could enable them to interpret human-defined rules, and thereby more easily navigate the human world.

Adapting to Dynamic Normative Contexts One powerful feature of RRC-aligned systems is the way that simple, heuristic decision-strategies are *connected* to more compute-intensive processes that come closer to the contractualist ideal (§3.3). Rules can be cached as outputs of universalization or virtual bargaining, for instance. Rules are often reliable and low-effort guides to action (hence their efficiency), but this is only true when certain important facts about the world remain constant. (An "emergency vehicles only" sign might be rendered moot once the emergency has been resolved.) This connection enables rules to be dynamically updated as the environment changes because an AI system would be able to fall back to the more flexible and context-specific contractualist processes that are sensitive to the changing environment (as well as the changing interests and values of the stakeholders). This may involve re-simulating what relevant stakeholders would agree to (or even eliciting additional human input) and re-codifying a novel rule that better fits the circumstances.

Assisting Human Moral Decision-Making This capacity to connect simple, heuristic rules with more resource-intensive processes also has the power to assist humans in their own moral decision-making. In creating law, humans often aim for something simple and easy to apply so that we can adapt our behavior accordingly, communicate them easily, and so on. But as a result, the law sometimes recommends something far short of the contractualist ideal—what would be agreed to if the affected parties could actually negotiate the specifics of the case. AI agents could enable us to exceed our existing resource rational compromises, and get closer to the ideal, by enabling us to apply more computation to solving the problem than we currently do.

Consider a town with a 10 PM noise ordinance. If, for a specific New Year's Eve party, all affected neighbors are not only invited but also consent to late-night festivities, this shared agreement becomes critical. Law enforcement, responding to the noise but unaware of this universal consent, would reasonably enforce the ordinance. However, possessing information about this unanimous agreement as well as an understanding of the function of the rule as an RRC approximation of that agreement (as an RRC-aligned system might) would justify permitting an exception to the rule. In this way, an RRC-aligned system could help humans more efficiently navigate social coordination problems that otherwise have human-imposed resource bottlenecks [10].

Reasonable Steerability There is recent interest in designing personal AI systems that are steerable to their principal's preferences or values [83]. However, steerability should have limits. An AI agent should not be able to severely or permanently harm others or prevent others' agents from achieving their goals. Agents, therefore, should be *reasonably* steerable. RRC provides a window into how this kind of bounded steerability could be operationalized. Given that agents will find themselves in novel contexts where it may be impossible to know precisely whether or not others would endorse or allow a particular action, RRC lays out a framework for approximating what is hypothetically mutually acceptable to the relevant stakeholders.

6 Future Directions

Process-level Supervision One possible implementation strategy parallels that of the recent "Deliberative Alignment" approach [28], though curating and including a range of RRC mechanism

thinking traces in the training set for process-level supervision [55]. This would involve providing the language model with a corpus of rules or policies, then training the model via supervised fine-tuning (SFT) to engage in explicit reasoning using a range of RRC strategies (e.g. as produced synthetically by chain-of-thought prompting) based on these guidelines.

Debate Protocols AI alignment employing debate protocols, as conceptualized by Irving et al. [38], typically involves multiple, distinct AI agents generating and defending divergent solution candidates in response to a user prompt. This process aims to enhance alignment by subjecting outputs of a powerful AI system to adversarial scrutiny by a similarly capable AI system before a final response is committed, reducing the cognitive burden on human evaluators who provide supervisory signals for AI training. RRC could leverage AI debate to instantiate contractualist mechanisms such as virtual bargaining, assigning each AI debater to represent different stakeholders, and thereby generating bargaining outcomes that a human committee could either accept or reject. The accepted outcomes could then be used to align AI systems via outcome-based supervision (effectively "caching" the contractualist outcome in the weights of the trained model) or process-based supervision (encouraging the trained model to reproduce human-endorsed contractualist reasoning).

Neuro-Symbolic Approaches An RRC approach is also well-suited for integration into neuro-symbolic architectures. Since symbolic representations enable both shared rules and individual agents' interests to be precisely encoded (e.g. as logical specifications in the former case [88, 63] and program-like utility functions in the latter [15, 95]), and can also ensure coherent probabilistic modeling of the world that agents interact in [14], mutual benefit can be formally specified, quantified, and verified. This in turn enables the design of algorithms that formally implement RRC mechanisms such as universalization or virtual bargaining — building on existing solvers from algorithmic game theory [72] — along with algorithms for sound meta-reasoning over which RRC mechanism to use. By combining these algorithms with LLMs that process natural language and multi-modal input into symbolic representations [90, 43, 94, 98], neuro-symbolic RRC architectures could unite both the open-endedness afforded by natural language with the reliability of symbolic systems.

RL for Mechanism Selection A central challenge for the RRC approach is how an agent would learn mechanism selection. This learning process could leverage recent advances in agentic architectures and reinforcement learning. The architectural foundation for RRC's meta-reasoning can be found in frameworks like ReAct, Reflexion, and especially Tree of Thoughts (ToT). These models operate on a meta-cognitive loop where the agent must choose its next operation - be it executing a tool, expanding a new line of reasoning, or reflecting on a past error. This choice architecture can serve as a substrate for RRC: the agent's decision on how wide or deep to expand its reasoning tree in ToT, for instance, directly operationalizes the choice between a fast, low-effort heuristic (a shallow search) and a costly, deliberative process (a deep search). To teach the model to make decisions using a resource-rational frame, the agent could be trained using Reinforcement Learning (RL) with a cost-aware reward function that values both accuracy and efficiency.

Data Collection for RRC-Alignment The implementation methods explored above, will likely require collecting large-scale and high-quality datasets. For instance, collecting examples of contractualist reasoning could be useful for training models to simulate bargains or any of their approximations (following the path laid out by Guan et al. [28]). This data could be collected from professional philosophers, gathered from high quality negotiation settings (such as in situations with professional mediators), or synthesized through guided chain of thought.

Another important avenue for data collection will be sourcing rules and norms that guide communities and the contractualist processes that generate and support them. Democratic processes that gain their legitimacy through a just political process are already in place to adjudicate between the pluralistic values of diverse citizens. Because RRC alignment takes explicit rules as a primary starting place for building an alignment target, these rules can be directly sourced from democratic processes and used in AI systems [46]. Another important source of data will be to gather norms that structure local communities (for instance, that are generated in community-based institutional settings). Online data collection from the public also offers a scalable method for informing RRC, enabling the elicitation of diverse perspectives on existing or proposed rules [19].

Acknowledgments

The authors would like to thank Gillian Hadfield, Nick Chater, Fiery Cushman, Max Kleiman-Weiner, Jon Gould, Becca Goldstein, Raphael Koster, Joe Edelman, and Ryan Lowe for their thoughtful contributions to the ideas presented in this paper. SL gratefully acknowledges the support of the Templeton World Charity Foundation (grant #TWCF-2023-32585).

Appendix

The appendix to this paper can be found at `https://arxiv.org/abs/2506.17434`.

References

[1] J Stacy Adams. Inequity in social exchange. In *Advances in experimental social psychology*, volume 2, pages 267–299. Elsevier, 1965.

[2] Elizabeth Anderson. *Value in ethics and economics*. Harvard University Press, 1995.

[3] Elizabeth Anderson. Symposium on amartya sen's philosophy: 2 unstrapping the straitjacket of'preference': a comment on amartya sen's contributions to philosophy and economics. *Economics and Philosophy*, 17(1):21, 2001.

[4] John Robert Anderson. *The adaptive character of thought*. Psychology Press, 1990.

[5] Jean-Baptiste André, Stephane Debove, Léo Fitouchi, and Nicolas Baumard. Moral cognition as a nash product maximizer: An evolutionary contractualist account of morality, May 2022. URL `psyarxiv.com/2hxgu`.

[6] Amanda Askell, Joe Carlsmith, Chris Olah, Jared Kaplan, Holden Karnofsky, and Anthropic. Claude's constitution, 2026. URL `https://www.anthropic.com/constitution`. Accessed: 2026-02-09.

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

[8] Federico Bianchi, Patrick John Chia, Mert Yuksekgonul, Jacopo Tagliabue, Dan Jurafsky, and James Zou. How well can llms negotiate? negotiationarena platform and analysis. In *Proceedings of the 41st International Conference on Machine Learning*, pages 3935–3951, 2024.

[9] Ken Binmore. *Natural justice*. Oxford university press, 2005.

[10] Justin B Bullock, Samuel Hammond, and Seb Krier. Agi, governments, and free societies. *arXiv preprint arXiv:2503.05710*, 2025.

[11] Alan Chan, Kevin Wei, Sihao Huang, Nitarshan Rajkumar, Elija Perrier, Seth Lazar, Gillian K Hadfield, and Markus Anderljung. Infrastructure for ai agents. *arXiv preprint arXiv:2501.10114*, 2025.

[12] Nick Chater and Mike Oaksford. Ten years of the rational analysis of cognition. *Trends in cognitive sciences*, 3(2):57–65, 1999.

[13] Nick Chater, Hossam Zeitoun, and Tigran Melkonyan. The paradox of social interaction: Shared intentionality, we-reasoning, and virtual bargaining. *Psychological Review*, 129(3):415–437, 2022.

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

[15] Guy Davidson, Graham Todd, Julian Togelius, Todd M Gureckis, and Brenden M Lake. Goals as reward-producing programs. *Nature Machine Intelligence*, 7(2):205–220, 2025.

[16] Oriel FeldmanHall, Tim Dalgleish, Davy Evans, Lauren Navrady, Ellen Tedeschi, and Dean Mobbs. Moral chivalry: Gender and harm sensitivity predict costly altruism. *Social psychological and personality science*, 7(6):542–551, 2016.

[17] James Fishkin, Nikhil Garg, Lodewijk Gelauff, Ashish Goel, Kamesh Munagala, Sukolsak Sakshuwong, Alice Siu, and Sravya Yandamuri. Deliberative democracy with the online deliberation platform. In *The 7th AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2019)*, pages 1–2, 2019.

[18] Bailey Flanigan, Paul Gölz, Anupam Gupta, Brett Hennig, and Ariel D Procaccia. Fair algorithms for selecting citizens' assemblies. *Nature*, 596(7873):548–552, 2021.

[19] Matija Franklin, Trisevgeni Papakonstantinou, Tianshu Chen, Carlos Fernandez-Basso, and David Lagnado. Blame attribution in human-ai and human-only systems: Crowdsourcing judgments from twitter. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 45, 2023.

[20] Robert L Frazier. Act utilitarianism and decision procedures. *Utilitas*, 6(1):43–53, 1994.

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

[22] Iason Gabriel and Geoff Keeling. A matter of principle? ai alignment as the fair treatment of claims. *Philosophical Studies*, pages 1–23, 2025.

[23] David Gauthier. *Morals by agreement*. Oxford University Press on Demand, 1986.

[24] Samuel J Gershman, Eric J Horvitz, and Joshua B Tenenbaum. Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, 349(6245): 273–278, 2015.

[25] Beth Goldberg, Diana Acosta-Navas, Michiel Bakker, Ian Beacock, Matt Botvinick, Prateek Buch, Renée DiResta, Nandika Donthi, Nathanael Fast, Ravi Iyer, et al. Ai and the future of digital public squares. *arXiv preprint arXiv:2412.09988*, 2024.

[26] Joshua David Greene. *Moral tribes: Emotion, reason, and the gap between us and them*. Penguin, 2014.

[27] Thomas L Griffiths, Falk Lieder, and Noah D Goodman. Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in cognitive science*, 7(2):217–229, 2015.

[28] Melody Y Guan, Manas Joglekar, Eric Wallace, Saachi Jain, Boaz Barak, Alec Heylar, Rachel Dias, Andrea Vallone, Hongyu Ren, Jason Wei, et al. Deliberative alignment: Reasoning enables safer language models. *arXiv preprint arXiv:2412.16339*, 2024.

[29] Julia Haas. Moral gridworlds: a theoretical proposal for modeling artificial moral cognition. *Minds and Machines*, 30(2):219–246, 2020.

[30] Julia Haas, Sophie Bridgers, Arianna Manzini, Benjamin Henke, Joshua May, Sydney Levine, Laura Weidinger, Murray Shanahan, Kristian Lum, Iason Gabriel, and William Isaac. A roadmap for evaluating moral competence in Large Language Models. *Nature*, 650(8102), Feb 2026. doi: 10.1038/s41586-025-10021-1. URL `https://www.nature.com/articles/s41586-025-10021-1`.

[31] Jürgen Habermas. *Moral consciousness and communicative action*. MIT press, 1990.

[32] Dylan Hadfield-Menell and Gillian K Hadfield. Incomplete contracting and ai alignment. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 417–422, 2019.

[33] Richard Mervyn Hare. *Moral thinking: Its levels, method, and point*. Oxford: Clarendon Press; New York: Oxford University Press, 1981.

[34] John C Harsanyi, Reinhard Selten, et al. A general theory of equilibrium selection in games. *MIT Press Books*, 1, 1988.

[35] Dan Hendrycks, Mantas Mazeika, Andy Zou, Sahil Patel, Christine Zhu, Jesus Navarro, Dawn Song, Bo Li, and Jacob Steinhardt. What would jiminy cricket do? towards agents that behave morally. *arXiv preprint arXiv:2110.13136*, 2021.

[36] Richard A Hirth, Michael E Chernew, Edward Miller, A Mark Fendrick, and William G Weissert. Willingness to pay for a quality-adjusted life year: in search of a standard. *Medical decision making*, 20(3):332–342, 2000.

[37] Eric J Horvitz. Reasoning about beliefs and actions under computational resource constraints. In *Proceedings of the Third Conference on Uncertainty in Artificial Intelligence*, pages 429–447, 1987.

[38] Geoffrey Irving, Paul Christiano, and Dario Amodei. Ai safety via debate. *arXiv preprint arXiv:1805.00899*, 2018.

[39] Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, et al. AI alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*, 2023.

[40] Ehud Kalai and Meir Smorodinsky. Other solutions to nash's bargaining problem. *Econometrica: Journal of the Econometric Society*, pages 513–518, 1975.

[41] Immanuel Kant. *Groundwork for the Metaphysics of Morals*. 1785.

[42] Oliver Klingefjord, Ryan Lowe, and Joe Edelman. What are human values, and how do we align ai to them? *arXiv preprint arXiv:2404.10636*, 2024.

[43] Joe Kwon, Sydney Levine, and Joshua B Tenenbaum. Neuro-symbolic models of human moral judgment: Llms as automatic feature extractors. 2023.

[44] Joseph Kwon, Zhi-Xuan Tan, Josh Tenenbaum, and Sydney Levine. When it's not out of line to get out of line: The rules of rule-breaking. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 45, 2023.

[45] Joseph Kwon, Josh Tenenbaum, and Sydney Levine. Neuro-symbolic models of human moral judgment. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46, 2024.

[46] Seth Lazar. Governing the algorithmic city. *Philosophy & Public Affairs*, 53(2):102–168, 2025.

[47] Seth Lazar and Lorenzo Manuali. Can llms advance democratic values? *arXiv preprint arXiv:2410.08418*, 2024.

[48] Arthur Le Pargneux, Nick Chater, and Hossam Zeitoun. Contractualist tendencies and reasoning in moral judgment and decision making. *Cognition*, 249:105838, 2024. ISSN 0010-0277. doi: https://doi.org/10.1016/j.cognition.2024.105838.

[49] Mark R Leary and Robin M Kowalski. Impression management: A literature review and two-component model. *Psychological bulletin*, 107(1):34, 1990.

[50] Sydney Levine, Max Kleiman-Weiner, Nick Chater, Fiery Cushman, and Joshua B Tenenbaum. The cognitive mechanisms of contractualist moral decision-making. In *CogSci*. Citeseer, 2018.

[51] Sydney Levine, Max Kleiman-Weiner, Laura Schulz, Joshua Tenenbaum, and Fiery Cushman. The logic of universalization guides moral judgment. In *Proceedings of the National Academy of Sciences*, 2020.

[52] Sydney Levine, Nick Chater, Joshua Tenenbaum, and Fiery Cushman. Resource-rational contractualism: A triple theory of moral cognition, 2024.

[53] Sydney Levine, Max Kleiman-Weiner, Nick Chater, Fiery Cushman, and Joshua B. Tenenbaum. When rules are over-ruled: Virtual bargaining as a contractualist method of moral judgment. *Cognition*, 250:105790, 2024. ISSN 0010-0277. doi: https://doi.org/10.1016/j.cognition.2024.105790.

[54] Falk Lieder and Thomas L Griffiths. Strategy selection as rational metareasoning. *Psychological Review*, 124(6):762, 2017.

[55] Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2023.

[56] Patricia L Lockwood, Miriam C Klein-Flügge, Ayat Abdurahman, and Molly J Crockett. Model-free decision making is prioritized when learning to avoid harming others. *Proceedings of the National Academy of Sciences*, 117(44):27719–27730, 2020.

[57] R Thomas McCoy, Shunyu Yao, Dan Friedman, Mathew D Hardy, and Thomas L Griffiths. Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, 121(41):e2322420121, 2024.

[58] Ryan M McManus, Max Kleiman-Weiner, and Liane Young. What we owe to family: The impact of special obligations on moral judgment. *Psychological Science*, 31(3):227–242, 2020.

[59] Jared Moore, Yejin Choi, and Sydney Levine. Intuitions of compromise: Utilitarianism vs. contractualism. *arXiv preprint arXiv:2410.05496*, 2024.

[60] Silen Naihin, David Atkinson, Marc Green, Merwane Hamadi, Craig Swift, Douglas Schonholtz, Adam Tauman Kalai, and David Bau. Testing language model agents safely in the wild. *CoRR*, 2023.

[61] John F Nash. The bargaining problem. *Econometrica: Journal of the econometric society*, pages 155–162, 1950.

[62] Richard Ngo, Lawrence Chan, and Sören Mindermann. The alignment problem from a deep learning perspective. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=fh8EYKFKns`.

[63] Ninell Oldenburg and Tan Zhi-Xuan. Learning and sustaining shared normative systems via bayesian rule induction in markov games. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, May 2024.

[64] OpenAI. Openai o1 system card. Technical report, OpenAI, Sep 2024. URL `https://openai.com/index/openai-o1-system-card/`. Accessed: 2026-02-10.

[65] L Owen and A Fischer. The cost-effectiveness of public health interventions examined by the national institute for health and care excellence from 2005 to 2018. *Public health*, 169:151–162, 2019.

[66] Derek Parfit. *On what matters: volume one*, volume 1. Oxford University Press, 2011.

[67] Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pages 1–18, 2022.

[68] Giorgio Piatti, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schölkopf, Mrinmaya Sachan, and Rada Mihalcea. Cooperate or collapse: Emergence of sustainability behaviors in a society of llm agents. *arXiv e-prints*, pages arXiv–2404, 2024.

[69] Peter Railton. Alienation, consequentialism, and the demands of morality. *Philosophy & Public Affairs*, pages 134–171, 1984.

[70] John Rawls. *A theory of justice*. Harvard university press, 1971.

[71] Min Reuchamps, Julien Vrydagh, and Yanina Welp. *De Gruyter handbook of citizens' assemblies*. De Gruyter, 2023.

[72] Tim Roughgarden. Algorithmic game theory. *Communications of the ACM*, 53(7):78–86, 2010.

[73] Jean-Jacques Rousseau. The social contract. *Londres*, 1762.

[74] Stuart J Russell. Rationality and intelligence. *Artificial intelligence*, 94(1-2):57–77, 1997.

[75] Stuart J Russell and Devika Subramanian. Provably bounded-optimal agents. *Journal of Artificial Intelligence Research*, 2:575–609, 1994.

[76] Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik Narasimhan, and Vishvak Murahari. Personagym: Evaluating persona agents and llms. *arXiv preprint arXiv:2407.18416*, 2024.

[77] Thomas Scanlon. *What we owe to each other*. Harvard University Press, 1998.

[78] Aaron Sell, Daniel Sznycer, Laith Al-Shawaf, Julian Lim, Andre Krauss, Aneta Feldman, Ruxandra Rascanu, Lawrence Sugiyama, Leda Cosmides, and John Tooby. The grammar of anger: Mapping the computational architecture of a recalibrational emotion. *Cognition*, 168: 110–128, 2017.

[79] Asankhaya Sharma. Autothink: efficient inference for reasoning llms. *Available at SSRN 5253327*, 2025.

[80] Alex Shaw. Beyond "to share or not to share" the impartiality account of fairness. *Current Directions in Psychological Science*, 22(5):413–417, 2013.

[81] Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. Large language model alignment: A survey. *arXiv preprint arXiv:2309.15025*, 2023.

[82] Haozhe Shi and Kun Niu. Enhancing persona consistency with large language models. In *Proceedings of the 2024 5th International Conference on Computing, Networks and Internet of Things*, pages 210–215, 2024.

[83] Taylor Sorensen, Pushkar Mishra, Roma Patel, Michael Henry Tessler, Michiel Bakker, Georgina Evans, Iason Gabriel, Noah Goodman, and Verena Rieser. Value profiles for encoding human variation. *arXiv preprint arXiv:2503.15484*, 2025.

[84] Wangtao Sun, Chenxiang Zhang, XueYou Zhang, Xuanqing Yu, Ziyang Huang, Pei Chen, Haotian Xu, Shizhu He, Jun Zhao, and Kang Liu. Beyond instruction following: Evaluating inferential rule following of large language models. *arXiv preprint arXiv:2407.08440*, 2024.

[85] Charles Taylor. What is agency. *Human agency and language: Philosophical papers*, 1:15–44, 1985.

[86] Michael Henry Tessler, Michiel A Bakker, Daniel Jarrett, Hannah Sheahan, Martin J Chadwick, Raphael Koster, Georgina Evans, Lucy Campbell-Gillingham, Tantum Collins, David C Parkes, et al. Ai can help humans find common ground in democratic deliberation. *Science*, 386(6719): eadq2852, 2024.

[87] Diego Trujillo, Mindy Zhang, Tan Zhi-Xuan, Joshua B. Tenenbaum, and Sydney Levine. Resource-rational virtual bargaining for moral judgment: Towards a probabilistic cognitive model. *TopiCS in Cognitive Science*, 2024.

[88] Georg Henrik Von Wright. On the logic of norms and actions. In *New studies in deontic logic: Norms, actions, and the foundations of ethics*, pages 3–35. Springer, 1981.

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

[90] Lionel Wong, Gabriel Grand, Alexander K Lew, Noah D Goodman, Vikash K Mansinghka, Jacob Andreas, and Joshua B Tenenbaum. From word models to world models: Translating from natural language to the probabilistic language of thought. *arXiv preprint arXiv:2306.12672*, 2023.

[91] Lionel Wong, Joseph Kwon, Junior Okorafoar, Mindy Zhang, Josh Tenenbaum, and Sydney Levine. What moral rules mean. in prep.

[92] Sarah A Wu, Xiang Ren, Tobias Gerstenberg, Yejin Choi, and Sydney Levine. Resource-rational moral judgment. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46, 2024.

[93] Qiuejie Xie, Qiming Feng, Tianqi Zhang, Qingqiu Li, Linyi Yang, Yuejie Zhang, Rui Feng, Liang He, Shang Gao, and Yue Zhang. Human simulacra: Benchmarking the personification of large language models. *International Conference on Learning Representations*, 2025.

[94] Lance Ying, Katherine M Collins, Megan Wei, Cedegao E Zhang, Tan Zhi-Xuan, Adrian Weller, Joshua B Tenenbaum, and Lionel Wong. The neuro-symbolic inverse planning engine (nipe): Modeling probabilistic social inferences from linguistic inputs. In *ICML 2023 Workshop on Theory of Mind in Communicating Agents*, July 2023.

[95] Wenhao Yu, Nimrod Gileadi, Chuyuan Fu, Sean Kirmani, Kuang-Huei Lee, Montserrat Gonzalez Arenas, Hao-Tien Lewis Chiang, Tom Erez, Leonard Hasenclever, Jan Humplik, et al. Language to rewards for robotic skill synthesis. In *Conference on Robot Learning*, pages 374–404. PMLR, 2023.

[96] Tan Zhi-Xuan. What should ai owe to us? accountable and aligned ai systems via contractualist ai alignment. *AI Alignment Forum*, 2022.

[97] Tan Zhi-Xuan, Micah Carroll, Matija Franklin, and Hal Ashton. Beyond preferences in ai alignment. *arXiv preprint arXiv:2408.16984*, 2024.

[98] Tan Zhi-Xuan, Lance Ying, Vikash Mansinghka, and Joshua B Tenenbaum. Pragmatic instruction following and goal assistance via cooperative language guided inverse plan search. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, May 2024.