Contractual AI: Toward More Aligned, Transparent, and Robust Dialogue Agents

Christopher J. Bates[1], **Ritwik Bose**[2,3], **Reagan G. Keeney**[2,3], **Vera A. Kazakova**[2,3]

[1] Harvard University, Dept. of Psychology
[2] Institute for Human and Machine Cognition
[3] Knox College
cjbates@g.harvard.edu, rbose@ihmc.org, rkeeney@ihmc.org, vkazakova@ihmc.org

Abstract

We present a new framework for AI alignment called *Contractual AI*, and apply it to the setting of dialogue agents chatting with humans. This framework incorporates and builds on previous approaches to alignment, such as *Constitutional AI*. We propose that fully aligned systems may need both a "think fast" and a "think slow" systems for approximating complex human judgements. Fast thinking (System 1) is computationally cheap but rigid and brittle in novel situations, while slow thinking (System 2) is more expensive but more flexible and robust. System 1 makes judgements by asking whether a rule or principle is violated. System 2 does the explicit reasoning that produces the rules, explicitly tallying costs and benefits for all stakeholders. Rule-based systems like *Constitutional AI* correspond roughly to System 1. Here, we implement a prototype of System 2, and lay out a road-map for enabling the system to make more thorough and accurate considerations for all stakeholder groups, including those underrepresented in the training data (e.g. racial minorities). For initial testing, we guided the decision process through the steps of: 1) identifying all stakeholders, 2) listing their individual concerns, 3) soliciting the projected opinions of various experts, and 4) combining the expert opinions into a final moral judgement. The resulting text was less generic, more aware of complex stakeholder needs, and ultimately more actionable.

Overview

If AI teammates are to be used at scale in a safe and ethical manner, it is critical that they be aligned to human values. In this work, we present a new cognitive architecture framework for AI alignment called *Contractual AI*, based on Contractualism, a promising unified model of human morality from cognitive science (Levine et al. 2023), and designed to meet technical needs such as reliability, transparency, and interpretability. While the framework is general, here we instantiate a version for the context of dialogue agents built on Transformer-based Large Language Models (LLM). Our approach builds directly on recent work arguing for *contractualism* in AI alignment (Awad et al. 2022; Jin et al. 2022; Kwon, Levine, and Tenenbaum 2023). Our *Contractual AI* is also complementary to previous rule-based approaches (Forbes et al. 2020; Solaiman and Dennison 2021), such as Anthropic's Constitutional AI (Bai et al. 2022), but addresses some of their key shortcomings: **transparency**, **insularity**, and **accuracy**.

Transparency: In Reinforcement Learning with Human Feedback (RLHF), inputs clearly label successful examples, but the resulting system's functionality remains "black-boxed", obscuring the system's reasoning, along with its biases and shortcomings. Our work "clear-boxes" the reasoning process by decomposing the problem of moral reasoning into separate modules, and leaving intermediate outputs as traces that can be inspected after the fact. This modular design also allows us to replace some computational steps that would otherwise be done by the LLM with simpler and more interpretable algorithms. An important design choice we make is to separate the step of identifying stakeholders and potential harms from the step of reasoning about those harms to come to a decision. Stratifying the problem allows agents (including humans) to employ more explicit thinking at every stage, including explicit agreements about how harms should be weighted (e.g. does harm to a child outweigh harm to an adult?), ultimately increasing transparency of the system's reasoning.

Insularity: current methods for AI alignment are created and overseen by small groups of people, with undisclosed positionality, who are not likely to represent all important stakeholder groups for each possible query, nor to pocess all the necessary expertise to meaningfully account for and address all stakeholder concerns. Furthermore, this structural gate-keeping reduces participation for minority stakeholder groups, contributing to inequitable practices. By contrast, the system we propose is explicitly designed to interface with advisory systems, allowing different communities to contribute specialized *advisor models* which can be deployed when the relevant community is identified as one of the stakeholders. For example, a dialogue system, on detecting that a user is a potential victim of domestic abuse and seeking advice, could deploy an advisor model trained to verifiably anticipate situational risks, overseen by certified human practitioners.

Accuracy: Current approaches do not fully capture the dynamics of human moral reasoning. Previous approaches to AI alignment have primarily relied on lists of rules and principles to nudge a dialogue agent toward more acceptable outputs. For example, Anthropic's Claude was aligned

using an actor-critic approach and Constitutional AI (Bai et al. 2022), wherein the critic was provided with a "constitution", i.e. a set of rules and principles to abide by (e.g. "do no harm"), and was asked to evaluate outputs from the actor (Claude) according to these guidelines. The actor then underwent further training based on these critiques. This approach has been very successful in aligning language models to human values. However, we argue that it cannot ever be fully aligned with humans, because it misses key aspects of human moral psychology, namely the way we seamless combine rule-based reasoning with partial-ordering of alternatives, as well as with rule-breaking when faced with extenuating circumstances.

Research in psychology highlights the flexibility of human morality (Levine et al. 2022; Kwon et al. 2023). We break or bend rules on a case-by-case basis, or even create new rules as needed, in predictable ways. For example, people generally follow the rule "No cutting in line", but most people would also permit a diabetic who urgently needs sugar to jump to the front of the line to buy a soda. An aligned AI system must understand not only the sets of general rules and principles that people adhere to, but also the calculus behind their exceptions. Here, we adopt "resource-rational contractualism" (Levine et al. 2023) as our guiding framework for approximating human values and reasoning in artificial systems, integrating both making/following rules and rule-breaking into one coherent system. It posits that people view actions as moral to the extent that they are consistent with a *virtual* contract, i.e. an agreement that society *would* come to if it were practical to do so formally. Under this view, rules are resource-rational—they are computationally (or cognitively) cheap and often serve as reasonable approximations to full contractual calculus.

Critical to our purposes, resource-rational contractualism can also be framed as a "dual process" theory, according to which people have two main reasoning systems at their disposal: one is cognitively cheaper but fallible (often labeled as "System 1" or "fast thinking") and the other is more expensive but generally more robust ("System 2" or "slow thinking"). Under this framing, offline rule-based fine-tuning approaches to alignment such as *Constitutional AI* constitute System 1. Our contribution is in providing the missing System 2 component in order to simulate human-like contractualism.

As AI capabilities continue to rapidly progress, governments and communities around the world must be able to certify systems' safety before they are deployed at scale. We believe that certification can only happen if AI systems 1) have behavior that is transparent and comprehensible, 2) allow for meaningful input from all stakeholder groups, and 3) are based on the most accurate and comprehensive theories of human morality at our disposal. The framework we present here aims to meet all of these criteria.

An Initial Design of Our System

Our system prototype employs an *actor-critic model*. The *actor* component generates a list of all stakeholders, potential responses to the user query, and predictions about harm to *stakeholders* for each response. These harms are then weighted according to human-aligned, contractualist principles. The *critic* component then generates *considerations* from the point of view of each stakeholder. These *considerations* are injected into the *actor*'s prompt and the process is repeated. There exist several approaches for composing contracts that satisfy all parties, although how best to implement these principles is an open area of investigation.

Council of Advisors A central challenge for System 2 is how to accurately identify and meaningfully incorporate all stakeholder considerations. For example, even if the system extracts the relevant stakeholders, it may not be able to correctly predict harms resulting from alternative judgements, especially for groups underrepresented or misrepresented in the training data. We argue that generic pretrained LLMs may not suffice, as generic reasoning can miss critical contextual clues, resulting in less accurate, less relevant, and/or less actionable judgments.

We propose crowd-sourcing expertise from advisory systems, which we call "council of advisors". Advisor models may be contributed by the community and can be constructed to promote the best interests of the stakeholder type they represent. *Advisors* can then be deployed at inference time whenever their expertise is identified as relevant, with each one independently analyzing the problem and providing a set of considerations, which are later taken into account by the *actor*. Concretely, an advisor could be a language model fine-tuned on stakeholder-specific data which takes the role of the *critic* in the above model. For our prototype, in lieu of available expert advising models, we instruct the core system to temporarily act in the capacity of each relevant advisor.

The Path Forward

To ensure computational viability, System 2 should be used judiciously. Humans are able to conserve cognitive effort by relying on System 1 whenever possible and only engaging System 2 in critical contexts. In order for our framework to be viable, substantial work may need to be devoted to emulating this ability, by designing reliable gating mechanisms which detect when to switch between systems. In addition, we can periodically consolidate knowledge from System 2 into System 1 by fine-tuning System 1 outputs based on those of System 2.

Our ongoing work is focused on refining several elements of the framework described above. We continue to refine the interaction of the *actor* and *critic* models, in particular identifying a variety interaction models which may be better suited to different types of problems. Further, careful specification of universal design principles for advisor models and identifying heuristics to dynamically select relevant advisor models is vital to build a robust ecosystem for the council of advisors. Finally, we are collecting situationally ambiguous queries on which to test the framework, focusing on problems whose solutions are contextually bound, rather than problems whose solutions are universally true and therefore easily trained on.

References

Awad, E.; Levine, S.; Loreggia, A.; Mattei, N.; Rahwan, I.; Rossi, F.; Talamadupula, K.; Tenenbaum, J.; and Kleiman-Weiner, M. 2022. When is it acceptable to break the rules? knowledge representation of moral judgement based on empirical data. *arXiv preprint arXiv:2201.07763*.

Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Forbes, M.; Hwang, J. D.; Shwartz, V.; Sap, M.; and Choi, Y. 2020. Social chemistry 101: Learning to reason about social and moral norms. *arXiv preprint arXiv:2011.00620*.

Jin, Z.; Levine, S.; Gonzalez Adauto, F.; Kamal, O.; Sap, M.; Sachan, M.; Mihalcea, R.; Tenenbaum, J.; and Schölkopf, B. 2022. When to make exceptions: Exploring language models as accounts of human moral judgment. *Advances in neural information processing systems*, 35: 28458–28473.

Kwon, J.; Levine, S.; and Tenenbaum, J. B. 2023. Neuro-Symbolic Models of Human Moral Judgment: LLMs as Automatic Feature Extractors. In *Workshop on Challenges of Deploying Generative AI: 40th International Conference on Machine Learning*.

Kwon, J.; Zhi-Xuan, T.; Tenenbaum, J.; and Levine, S. 2023. When it is not out of line to get out of line: The role of universalization and outcome-based reasoning in rule-breaking judgments. *PsyArXiv. osf.io/preprints/psyarxiv/n8bjr*.

Levine, S.; Chater, N.; Tenenbaum, J.; and Cushman, F. A. 2023. Resource-rational contractualism: A triple theory of moral cognition. *PsyArXiv. osf.io/preprints/psyarxiv/p48t7*.

Levine, S.; Kleiman-Weiner, M.; Chater, N.; Cushman, F. A.; and Tenenbaum, J. 2022. When rules are over-ruled: Virtual bargaining as a contractualist method of moral judgment. *PsyArXiv. osf.io/preprints/psyarxiv/k5pu8*.

Solaiman, I.; and Dennison, C. 2021. Process for adapting language models to society (palms) with values-targeted datasets. *Advances in Neural Information Processing Systems*, 34: 5861–5873.