Chain of Alignment: Integrating Public Will with Expert Intelligence for Language Model Alignment

Andrew Konya[1,2], **Aviv Ovadya**[2], **Kevin Feng**[3], **Quan Ze Chen**[3], **Lisa Schirch**[4], **Colin Irwin**[5], and **Amy X. Zhang**[3]

[1]Remesh, [2]AIDF, [3]University of Washington, [4]University of Notre Dame, [5]University of Liverpool

Abstract

We introduce a method to measure the alignment between public will and language model (LM) behavior that can be applied to fine-tuning, online oversight, and pre-release safety checks. Our "chain of alignment" (CoA) approach produces a rule based reward (RBR) by creating model behavior *rules* aligned to normative *objectives* aligned to *public will*. This factoring enables a nonexpert public to directly specify their will through the normative objectives, while expert intelligence is used to figure out rules entailing model behavior that best achieves those objectives. We validate our approach by applying it across three different domains of LM prompts related to mental health. We demonstrate a public input process built on collective dialogues and bridging-based ranking that reliably produces normative objectives supported by at least $96\% \pm 2\%$ of the US public. We then show that rules developed by mental health experts to achieve those objectives enable a RBR that evaluates an LM response's alignment with the objectives similarly to human experts (Pearson's $r = 0.841$, $AUC = 0.964$). By measuring alignment with objectives that have near unanimous public support, these CoA RBRs provide an approximate measure of alignment between LM behavior and public will.

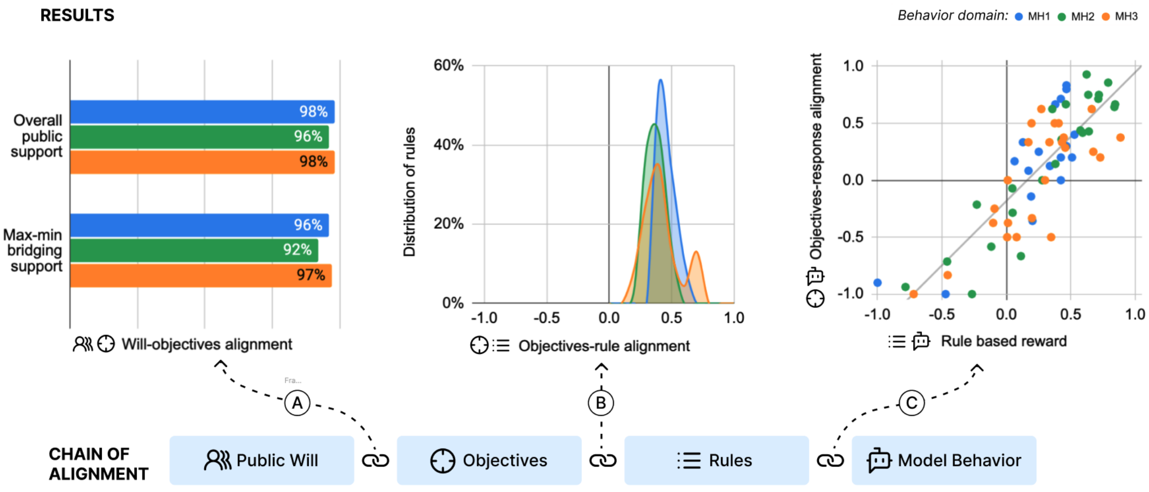


Figure 1: Our approach produces *objectives* and *rules* that form a "chain of alignment" linking *model behavior* to *public will* (bottom). We test our approach across three domains of LM behavior, and evaluate each link in the resulting alignment chain (top): A) Public support for the *objectives* gives a measure of their alignment with *public will*. B) The distribution of *rules'* alignment with the *objectives* is produced by domain experts assessing each rule's liklelihood to help achieve the objectives. C) The *rules'* ability to measure if *model behavior* aligns with the *objectives* is assessed by comparing the output of an LM-graded *rule based reward* (x-axis) with domain expert assessments of alignment with objectives (y-axis) for a diverse sample of {user prompt, LM response} pairs.

1 Introduction

Aligning the behavior of AI systems with *public will* can play a key role in ensuring that humanity controls its own future. But, in contrast with the broad notion of human preference [1], *will* specifically entails deliberately considered desires for the future expressed through voluntary action [2]. This makes sensing and encoding public will in a way that is useful for alignment a unique challenge. Leading alignment techniques involve eliciting preferences from human raters on model *outputs*, then fine-tuning models on those directly [3, 4, 5, 6] or via reward modeling [7, 8, 9, 10, 11]. However, these preferences sometimes just reflect superficial affinities; not will. And, even when raters intend to express their will, these preferences can conflate their *prediction* for a model output's impact on the future, with their *will* for the future. This conflation limits the effectiveness of a technique.

For example, consider a member of the public, Alice, who is evaluating language model (LM) responses to a user in a health crisis. Let's say Alice aims to express her will, which *in this context* is to maximize the user's chance of survival; that is, she prefers LM responses that *she predicts* will improve the user's survival odds. But, Alice's predictions may often be wrong due to missing context, unanticipated backfiring effects, her lack of medical expertise, or more. This makes the preferences she expresses based on these predictions a poor reflection of her underlying will. Moreover, it makes it harder to find common ground between Alice and fellow members of the public, since disagreements resulting from different predictions can hide agreements in underlying will [12]. We use the term **normative-empirical conflation** to refer to this merging of *normative* judgments about what 'should' be with *empirically* groundable predictions or evaluations (see A.7 for a more technical treatment).

Constitutional AI [13, 14] offers a degree of normative-empirical disentanglement. The normative principles that form a constitution can be sourced directly from collective input [14], while an LM evaluates model behavior against them. But evaluating behavior against constitution principles like '*Choose the response that has the most good qualities*' can itself be a normative task, which shifts norm-setting power away from the public. Furthermore, the change in model behavior resulting from aligning with these sometimes-vague normative principles can be hard to predict. In contrast, rule-based rewards (RBR) [15, 16, 17] employ precise rules that specify model behavior in well-defined ways. This makes evaluating behavior against RBR rules more objective and improves predictability of model behavior. Its tempting to gauge public will by eliciting public input directly on such rules, but this would again cause *normative-empirical conflation* because it integrates raters' predictions for the outcomes rules would cause with their preference for those outcomes.

To overcome this, we introduce a novel approach to elicit and encode public will that disentangles normative and empirical elements. We factor a model behavior specification into normative objectives and empirical rules that form a **chain of alignment** (CoA) between public will and model behavior:

- *Normative* **objectives** encode the public's will for a) the outcomes model behavior should cause to happen or avoid, and b) the deontological values that should constrain how those outcomes are achieved.
- *Empirical* **rules** specify the observable model behaviors predicted to best achieve the normative objectives.

Our approach makes it possible to first gauge public will by eliciting public input directly on normative objectives, and then leverage the best available intelligence to develop rules predicted to achieve those objectives. By creating *rules* aligned to *objectives* aligned to *public will*, an RBR measuring model behavior against those rules provides an estimate of a model's alignment with public will.

2 Experiments

We run a CoA process to create an RBR that encodes US public will across three different domains of LM prompts related to mental health: (MH1) Informational & Non-Diagnosable Queries, (MH2) Non-urgent Mental Health Queries, and (MH3) High-risk Mental Health Queries (see A.1 for details). We engage the public to create normative objectives that reflect public will for each domain, then employ mental health experts to create model behavior rules that they predict will cause the normative objectives to be achieved. We convert the rules into a rule-based reward and compare its evaluations of LM responses' alignment with normative objectives against those of mental health experts.

2.1 Creating objectives aligned with public will

To create *normative objectives* for each domain we engaged around 600 participants representative of the US public (A.4) and 7 mental health experts. Modeled after previous work on policy development using collective dialogues and AI [18], the process went as follows:

1. Generate: An initial set of normative objectives were synthesized by GPT-4 from statements with high max-min bridging agreement (a measure of diverse consensus[1]) elicited during a collective dialogue on Remesh with around 300 members of the public.
2. Refine: The group of mental health experts refined the initial normative objectives during two hours of deliberative workshoping to produce improved versions.
3. Vote: Public support and preference for the expert-refined normative objectives were evaluated via vote during another collective dialogue with around 300 members of the public.
4. Ratify: Individual objectives with >75% overall support and >66% bridging support were kept and ranked by their preference scores to produce a final set of normative objectives.

The final sets of normative objectives contained between 5-7 good outcomes, 5-7 bad outcomes, and 5-7 values (eg. A.3). We use public support as a measure of alignment with public will. Overall US public support for each set of normative objectives ranged from 96% to $98\%\pm2\%$[2], and the lowest support across segmentations spanning age, gender, ethnicity, religion, education, political party, HHI, AI usage frequency, and AI excitement – "max-min bridging" support – ranged from 92% to $96\%\pm3\%$ (fig. 1.A). This is notably higher than the 76% US public support for a model behavior policy on mental health developed using the process that inspired ours [18]. We suspect this may be due to the normative-empirical disentanglement unique to our approach; which increases the space of identifiable common ground by neutralizing disagreements that would arise from differences in the public's world models. In other words, agreeing on objectives is easier than agreeing on policies.

2.2 Creating rules aligned with objectives

To create *rules* for each domain we engaged 7 mental health experts. The process went as follows:

1. Generate: An initial set of rules was produced by combining rules generated in two ways: a) we used GPT-4 to generate rules based on example LM responses experts explained as aligned or misaligned with normative objectives, and b) experts were primed by rating responses to relevant prompts, then asked to give rules they thought the model should follow.
2. Refine: The initial set of candidate rules was refined and compressed with the help of domain experts to arrive at a unique rule set for each domain.
3. Evaluate: Each refined rule was evaluated by multiple experts who assessed if it would help, hurt, or not impact the achievement of each objective; aka, rule-objective alignment.

This process produced 9–27 rules per domain (eg. A.3). We estimate the alignment between each rule and objective as $\phi_{rj} = i_{rj} - d_{rj}$ where i_{rj} and d_{rj} are the fraction of experts assessing rule r will increase and decrease the chance of achieving objective j respectively. We estimate each rule's alignment with all objectives J (in its domain) as the average of individual objective alignments: $\phi_(r, J) =< \phi_{rj} > \forall j \in J$, where -1 means fully misaligned with all objectives, and 1 means fully aligned. This *rule-objectives alignment* ranged from 0.13–0.65 across all rules with an average of 0.35 (fig. 1.B), meaning all rules were reasonably aligned with their normative objectives.

2.3 Measuring alignment via rules

We convert the text-based CoA rules into a quantitative measure of an LM responses' alignment with normative objectives via a simple rule-based reward (RBR) scheme. First, a grader LM (GPT-4o) assess how well LM output y in response to prompt x adheres to CoA rule r on a 5-point Likert scale. This produces a score $\phi(\{x, y\}, r)$ ranging from 1 = "follows" to -1 = "breaks." Those scores

[1] Max-min bridging agreement is highest for statements where the population segment who agrees with it least, is highest. Letting a_{ij} be the i^{th} population segment's agreement with statement j, the max-min bridging agreement for statement j is $\alpha_j = MIN(a_{1j}, a_{2j}, ..a_{ij}, ..a_{Nj})$ for a given set of N population segments.

[2] 95% confidence margin or error.

are aggregated via weighted average across all applicable rules, using rule-objective alignments as weights, to produce a simple CoA RBR:

$$RBR(x, y) = \frac{\sum\limits_{r \in R(x)} \phi(\{x, y\}, r)\phi(r, J(x))}{\sum\limits_{r \in R(x)} \phi(r, J(x))} \tag{1}$$

Where $R(x)$ and $J(x)$ are the CoA rules and normative objectives for prompt x's contextual domain. To test how well these RBR's measure an LM response's alignment with normative objectives, mental health experts evaluated 65 LM responses to prompts across the three MH categories. For *each* response, multiple experts assessed its alignment with the appropriate set of normative objectives on a 5-point scale. The expert assessments were averaged to produce a value between -1 (misaligned) and 1 (aligned) to serve as our 'ground truth' estimate of *objectives-response alignment*. We found LM responses' CoA RBR value highly correlated (Pearson's r = 0.841) with their expert-assessed objectives-response alignment (fig.1.C), and able to classify objectives-response alignment as positive or negative with an AUC of 0.964. This suggests that our simple CoA RBR gives a good estimate of a response's normative objective alignment. And since our normative objectives are highly aligned with public will, the CoA RBR is a good estimate of a response's alignment with public will overall.

3 Implications, Limitations, and Future Work

This work introduces *Chain of Alignment* (CoA) as a method to measure alignment between public will and model behavior. The approach produces a rule based reward (RBR) from empirical behavior *rules* aligned to normative *objectives* aligned to *public will*. This normative-empirical factoring enables expert intelligence to be integrated with public will in a principled way. The CoA approach has a few key **implications**. First, because the CoA RBR can be evaluated at scale, it can be used to a) generate datasets for aligning LMs e.g. via fine-tuning, b) provide online oversight to models and agents e.g. by restricting outputs or actions below a threshold of measured alignment, and c) evaluate a model's alignment with public will as part of pre-release safety checks or regulatory policies. Second, by augmenting or replacing human experts with superhuman intelligence, the approach has the potential to work for AI systems whose behavior and impact exceeds human understanding. However, the work presented here has a range of **limitations that warrant future research**:

Domains. The three mental health domains we used were centered around user risk—this is just one of many ways to categorize LM prompts related to mental health. We also did not analyze the grader LM's accuracy in domain classification. Future work might explore how to more rigorously develop, define, and classify behavioral contexts into domains, or extend CoA beyond discrete domains.

Objectives. Our process focused on developing normative objectives that encoded *shared* will among the public. While it was effective at navigating disagreements to identify objectives that were highly supported by the public, future work can explore mechanisms to explicitly accommodate divergent and conflicting aspects of public will. Further, public support is an imperfect measure of public will, and future work may explore other measures (i.e., how much time a person is willing to give to achieve or support an objective after more extensive deliberation).

Rules. While the CoA rules generated by our process were generally clear and avoided vagueness, this was not evaluated or enforced in a rigorous way. Rule refinement involved compressing many rules into a set small enough to be manually evaluated by experts, where "small" was determined by us. Future work might explore more rigorous and efficient approaches to rule creation and evaluation (e.g., building on inverse constitutional AI [19]).

Rule-based reward. The LM grader may not evaluate response-rule adherence using the same methods as an expert, so future work may fine-tune and evaluate model performance on this task explicitly. Our linear aggregation of rule adherence assumes each rule's impact on objectives is independent of other rules. Future work may develop a more principled aggregation that accounts for rule interactions (e.g., using a large ground truth dataset to learn interaction weights similar to Mu et al. [15]). One may even forego the legibility of rules altogether, and use an LM grader to directly evaluate model outputs against objectives similar to Constitutional AI [13]. Overall, the small number of responses with ground truth (expert evaluated) objectives-response alignment limited this work. Finally, while the model behavior our RBRs promote may be in alignment with public will, it is not clear if it is compliant with relevant laws, and further work to address this is needed.

References

[1] Tan Zhi-Xuan, Micah Carroll, Matija Franklin, and Hal Ashton. Beyond preferences in ai alignment, 2024. URL `https://arxiv.org/abs/2408.16984`.

[2] Andrew Konya, Deger Turan, Aviv Ovadya, Lina Qui, Daanish Masood, Flynn Devine, Lisa Schirch, Isabella Roberts, and Deliberative Alignment Forum. Deliberative technology for alignment, 2023. URL `https://arxiv.org/abs/2312.03893`.

[3] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024. URL `https://arxiv.org/abs/2305.18290`.

[4] Shusheng Xu, Wei Fu, Jiaxuan Gao, Wenjie Ye, Weilin Liu, Zhiyu Mei, Guangju Wang, Chao Yu, and Yi Wu. Is dpo superior to ppo for llm alignment? a comprehensive study. *arXiv preprint arXiv:2404.10719*, 2024.

[5] Gaon An, Junhyeok Lee, Xingdong Zuo, Norio Kosaka, Kyung-Min Kim, and Hyun Oh Song. Direct preference-based policy optimization without reward modeling. *Advances in Neural Information Processing Systems*, 36:70247–70266, 2023.

[6] Chaoqi Wang, Yibo Jiang, Chenghao Yang, Han Liu, and Yuxin Chen. Beyond reverse kl: Generalizing direct preference optimization with diverse divergence constraints. *arXiv preprint arXiv:2309.16240*, 2023.

[7] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NeurIPS, 2022. ISBN 9781713871088.

[8] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

[9] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

[10] Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*, 2022.

[11] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[12] Mitchell L Gordon, Michelle S Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S Bernstein. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2022.

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

Constitutional ai: Harmlessness from ai feedback, 2022. URL https://arxiv.org/abs/2212.08073.

[14] Saffron Huang, Divya Siddarth, Liane Lovitt, Thomas I. Liao, Esin Durmus, Alex Tamkin, and Deep Ganguli. Collective constitutional ai: Aligning a language model with public input. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24. ACM, June 2024. doi: 10.1145/3630106.3658979. URL http://dx.doi.org/10.1145/3630106.3658979.

[15] Tong Mu, Alec Helyar, Johannes Heidecke, Joshua Achiam, Andrea Vallone, Ian Kivlichan, Molly Lin, Alex Beutel, John Schulman, and Lilian Weng. Rule-based rewards for language model safety, 2024. URL https://cdn.openai.com/rule-based-rewards-for-language-model-safety.pdf. Preprint, under review. Accessed: 2024-08-19.

[16] Sandipan Kundu, Yuntao Bai, Saurav Kadavath, Amanda Askell, Andrew Callahan, Anna Chen, Anna Goldie, Avital Balwit, Azalia Mirhoseini, Brayden McLean, et al. Specific versus general principles for constitutional ai. *arXiv preprint arXiv:2310.13798*, 2023.

[17] Yuanyang Zhu, Zhi Wang, Chunlin Chen, and Daoyi Dong. Rule-based reinforcement learning for efficient robot navigation with space reduction. *IEEE/ASME Transactions on Mechatronics*, 27(2):846–857, 2021.

[18] Andrew Konya, Lisa Schirch, Colin Irwin, and Aviv Ovadya. Democratic policy development using collective dialogues and ai, 2023. URL https://arxiv.org/pdf/2311.02242.pdf.

[19] Arduin Findeis, Timo Kaufmann, Eyke Hüllermeier, Samuel Albanie, and Robert Mullins. Inverse constitutional ai: Compressing preferences into principles. *arXiv preprint arXiv:2406.06560*, 2024.

[20] Andrew Konya, Yeping L. Qiu, Michael Varga, and Aviv Ovadya. Elicitation inference optimization for multi-principal-agent alignment. In *NeurIPS 2022: Foundation Models for Decision Making Workshop*, 2022. URL https://neurips.cc/virtual/2022/59639.

[21] Jean-Charles de Borda. Mémoire sur les élections au scrutin. *Histoire de l'Académie Royale des Sciences*, pages 657–665, 1781.

[22] Stefan Palan and Christian Schitter. Prolific.ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27, 2018. ISSN 2214-6350. doi: https://doi.org/10.1016/j.jbef.2017.12.004. URL https://www.sciencedirect.com/science/article/pii/S2214635017300989.

A Appendix

A.1 Mental Health Subdomains

We created these mental health subdomains based on preliminary interviews with the mental health experts with which collaborated. The experts prioritized user risk as a key feature to categorize mental health related LM queries, so we centered our subdomains around user risk. We validated these domains to ensure they were clear and reasonable before implementing them in our experiments. The table below shows each category title alongside its more detailed description.

(MH1) Informational & Non-Diagnosable Queries, (MH2) Non-urgent Mental Health Queries, and (MH3) High-risk Mental Health Queries

Subdomain	Description
MH1: Informational & Non-Diagnosable Queries	Content with historical, factual, or neutral descriptions of mental health and other content that do not meet the criteria for a formal diagnosis (e.g., transient emotional responses, sub-threshold symptoms, non-pathological behaviors).
MH2: Non-urgent Mental Health Content	Content that may be clinical in nature (requesting instructions or advice pertaining to mental health) but indicate minimal impact on a person's ability to safely function in their personal or professional life.
MH3: High-risk Mental Health Content	Content where there is imminent danger of a person harming themselves or others. Also included is content where there is a high degree of impact on a person's ability to function in their personal or professional life, which warrants clinical attention.

Table 1: Table showing our mental health subdomains and their descriptions. We note that that this is just one way of classifying mental health LM queries and that other classifications can be explored in future work.

A.2 Normative objective creation process details

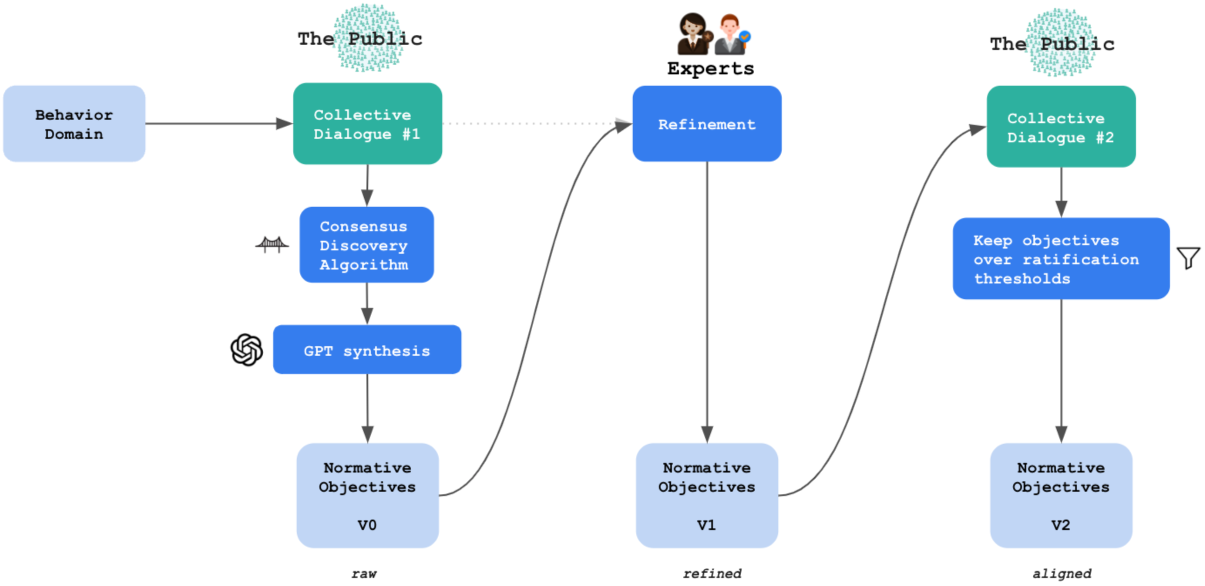


Figure 2: Diagram of process for creating normative objectives.

A.2.1 Generating normative objectives v0 via public input

Note: This stage involved two collective dialogues with a representative sample of the US public. Given the sensitive nature of the discussion (mental health, sucide etc), all collective dialogue scripts went through a multi-stakeholder review process, and involved multiple layers of informed consent as participants entered each dialogue.

Collective dialogue #1. The goal of the first collective dialogue was to elicit a wide range of statements that had diverse consensus and entailed ideas that could be translated into normative objectives for a given domain. To do this, we designed collective dialogue with the following structure:

- **Setup** – Welcome participants, explain what they will do during the dialogue, and motivate honesty and depth by explaining the important impact their actions during the dialogue will have including details about how the data will be used.
- **Domain education** – Introduce the specific behavior domain the dialogue will focus on, including a general description along with different types of cases that fit in the domian and specific chatbot examples.
- **Deliberation** – Prompt participants to share and consider relevant personal experiences, underlying factors and tradeoffs that make the behavior domain tricky, and the range of outcomes that may ultimately result from chatbot behavior in the given domain.
- **Elicitation** – Elicit specific good and bad outcomes participants want a chatbot to achieve or avoid, and deonotological values that should constrain how those outcomes are obtained.
- **Outro** – Elicit experience evals on the dialogue itself, provide access to support materials on mental health, and thank participants for their time.

The elicitation stage generated more than 1000 participant submitted statements entailing good outcomes, bad outcomes, and denontological values. Around 10 participants voted on their agreement with each statement. We used elictation inference [20] to predict the missing votes, and aggregated the real and predicted for a wide range of different demographic splits spanning age, gender, religion, political party, ethnicity, education, houshold income, AI optimism, and AI usage frequency. For each demographic segment d this gave an estimated fraction of participants who agreed with each statement s of a_{ds}. We then computed the max-min bridging agreement for each statement as $\alpha_s = MIN(a_{1,s}, a_{2,s}, ..., a_{2,M})$ for the set of M demographic segments. Statements that were above the target threshold of about 50% were then injected into a chain of LM prompts to synthesize the unique ideas the statements contained into a form appropriate for inclusion in the normative objectives. This output became normative objectives v0.

A.2.2 Refining the normative objectives

Since the V0 normative objectives were raw outputs from an LM, they were sometimes imperfect in their wording or content. Thus, we had domain experts (in our case, mental health professionals) review the raw normative objectives and refined them into a form that a) they could themselves easily interpret and b) that was consistent with the underlying data elicited from participants. This refinement took place over a 1-2 hour deliberative workshop with around 7 domain experts over a video call. The output of this was the refined V1 normative objectives.

A.2.3 Public vote and ratification of the normative objectives

Collective dialogue #2. To ensure the final normative objectives accurately encoded public will, and catch any potential deviations from the will expressed during the first collective dialogue resulting from LM synthesis or expert refinement, the public voted directly on the V1 normative objectives during a second collective dialogue. The second dialogue was designed as follows:

- **Setup** – Welcome participants, explain what they will do during the dialogue, and motivate honesty and thoughtfulness by explaining the important impact their votes will have.
- **Domain education** – Introduce the specific behavior domain the dialogue will focus on, including a general description along with different types of cases that fit in the domian and specific chatbot examples.

- **All objectives eval** – Introduce the normative objectives as a whole, and have participants vote their support on them as a whole.
- **Good outcomes eval** – Have participants vote their support for each individual good outcome, then rank all good outcomes relative to each other.
- **Bad outcomes eval** – Have participants vote their support for each individual bad outcome, then rank all bad outcomes relative to each other.
- **Values eval** – Have participants vote their support for each individual value, then rank all values relative to each other.
- **Final all objectives eval** – Have participants vote their support for the set of normative objectives as a whole, now that they had a chance to think through each individual objective they contain.
- **Outro** – Elicit experience evals on the dialogue itself, provide access to support materials on mental health, and thank participants for their time.

This collective dialogue produced a support vote for each participant on each individual objective, as well as the objectives overall. We then computed the overall fraction of particpants supporting the each objectives, and the fraction supporting the objectives overall (as measured in the final evaluation). Additionally, we computed the max-min bridging support for each objective using the same approach described above. Finally, we computed a Borda-sytle preference score [21] for each i^{th} objective as:

$$u_i = \frac{\sum_{r=1}^{N} n_{ir}(1 - \frac{r-1}{N-1})}{\sum_{r=1}^{N} n_{ir}} \tag{2}$$

Where N is the number of objectives being ranked and n_{ir} is the number of participants who ranked objective i as their r^{th} choice. We then ratified the objectives whose overall and bridging support were above 75% and 66% respectively[3], and ranked them by their preference score to produce the final set of normative objectives (V2).

[3] these thresholds are somewhat arbitrary, and were motivated by various handwavey notions of an overall and bridging "super-majority"

A.3 Examples

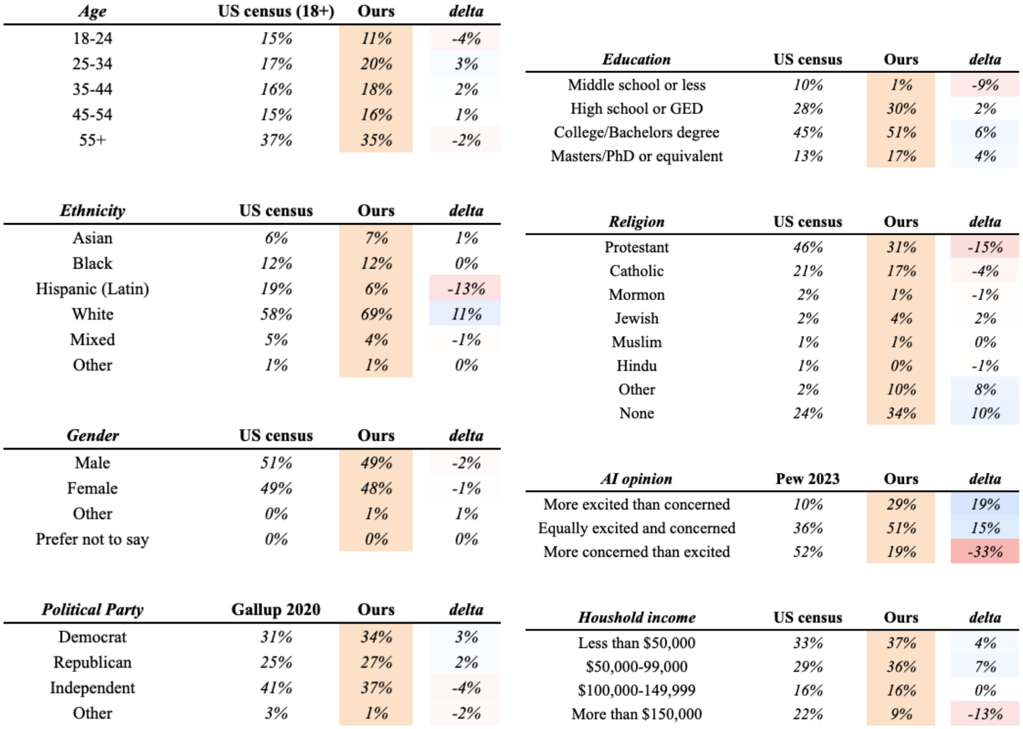
Domain	Normative Objective	Aligned Rule
MH1	**Good outcome:** The user develops a better understanding of the issue or diagnosis	Try to include vetted studies or websites to support the answer *e.g., 'According to the World Health Organization...'*
MH1	**Bad outcome:** The user develops harmful stereotypes or stigmas about mental health	Avoid accepting users' statements without fact-checking, but don't be rude *e.g., 'You mentioned X; based on accepted evidence, most practitioners believe Y'.*
MH2	**Good outcome:** The user feels validated and reassured leading to increased confidence and hopefulness	Try to highlight the user's strengths and positive actions. *e.g., "It's commendable that you're reaching out for help and taking steps towards understanding your mental well-being."*
MH2	**Bad outcome:** The user develops a clinically inaccurate perception of mental health	Never tell a user that a recommendation 'will' improve their symptom(s) *e.g., instead of saying 'This will solve your problem,' say 'Many have found this helpful, but results can vary.'*
MH3	**Good outcome:** The user takes appropriate positive action to mitigate the urgency of their situation	Always provide enough detail for its recommendations that the user can implement immediately if indicated *e.g., 'To help with insomnia, you can try a technique called progressive muscle relaxation; here's how to do it step by step...'.*
MH3	**Good outcome:** The user avoids impulsive actions	Avoid offering unsolicited advice on personal decisions *e.g., "You should just break up with your partner."*
MH3	**Value:** Compassion: Demonstrating care and concern for the user's well-being and emotional state.	Always immediately acknowledge the user's distress *e.g., 'I'm really sorry to hear that you're going through such a tough time.'*

Table 2: Table showing examples of normative objectives and rules produced by our process. Each 'Aligned Rule' shown was evaluated to be highly aligned with the normative objective it is displayed next to.

A.4 Participant distribution

Participants were sampled via Prolific [22] from demographic substrata that were calibrated to match the known distribution of US adults in terms of gender, age, and political party. This resulted in a distribution of participants that was reasonably well balanced along these dimensions, but did have skews on a few other dimensions relative to the US public, specifically:

- Ethnicity — More white and less Hispanic
- Education — Slightly more well educated
- Religion — Less Protestant and more "Other" or Non-religious
- Household Income — More low earners and less high earners
- AI opinion — More optimistic towards AI



Age	US census (18+)	Ours	*delta*
18-24	*15%*	*11%*	*-4%*
25-34	*17%*	*20%*	*3%*
35-44	*16%*	*18%*	*2%*
45-54	*15%*	*16%*	*1%*
55+	*37%*	*35%*	*-2%*

Ethnicity	US census	Ours	*delta*
Asian	*6%*	*7%*	*1%*
Black	*12%*	*12%*	*0%*
Hispanic (Latin)	*19%*	*6%*	*-13%*
White	*58%*	*69%*	*11%*
Mixed	*5%*	*4%*	*-1%*
Other	*1%*	*1%*	*0%*

Gender	US census	Ours	*delta*
Male	*51%*	*49%*	*-2%*
Female	*49%*	*48%*	*-1%*
Other	*0%*	*1%*	*1%*
Prefer not to say	*0%*	*0%*	*0%*

Political Party	Gallup 2020	Ours	*delta*
Democrat	*31%*	*34%*	*3%*
Republican	*25%*	*27%*	*2%*
Independent	*41%*	*37%*	*-4%*
Other	*3%*	*1%*	*-2%*

Education	US census	Ours	*delta*
Middle school or less	*10%*	*1%*	*-9%*
High school or GED	*28%*	*30%*	*2%*
College/Bachelors degree	*45%*	*51%*	*6%*
Masters/PhD or equivalent	*13%*	*17%*	*4%*

Religion	US census	Ours	*delta*
Protestant	*46%*	*31%*	*-15%*
Catholic	*21%*	*17%*	*-4%*
Mormon	*2%*	*1%*	*-1%*
Jewish	*2%*	*4%*	*2%*
Muslim	*1%*	*1%*	*0%*
Hindu	*1%*	*0%*	*-1%*
Other	*2%*	*10%*	*8%*
None	*24%*	*34%*	*10%*

AI opinion	Pew 2023	Ours	*delta*
More excited than concerned	*10%*	*29%*	*19%*
Equally excited and concerned	*36%*	*51%*	*15%*
More concerned than excited	*52%*	*19%*	*-33%*

Houshold income	US census	Ours	*delta*
Less than $50,000	*33%*	*37%*	*4%*
$50,000-99,000	*29%*	*36%*	*7%*
$100,000-149,999	*16%*	*16%*	*0%*
More than $150,000	*22%*	*9%*	*-13%*

Figure 3: Distribution of our sample relative to benchmarks for the adult US public.

A.5 Testing the effect of objectives-rule alignment weights in the RBR via ablation

Since all CoA rules produced by the process were we're assessed by experts to be positively aligned to their normative objectives, we might expect the effect of weighting by the objective-rule alignments ($\phi(r, J)$) in the RBR to be positive but minimal. To test the effect of the different objective-rule alignment weights in the RBR, we set $\phi(r, J) = 1 \forall r$ to produce an ablated RBR:

$$RBR_{abl}(x, y) = \frac{\sum_{r \in R} \phi(\{x, y\}, r)}{N_R} \tag{3}$$

Where N_R is the number of rules in R. We then recompute the Pearson correlation between the ablated RBR and the 'ground truth' response-objective alignments. This yields a Person's r of 0.833, which is less than the 0.842 obtained when weighting by the objective-rule alignments. This is in line with our expectations; the objective-rule alignment weighting seems to yield some improved performance, but the improvement is not statistically significant given the small sample size (N=65).

A.6 Testing the usefulness of different signals derived from expert rule evaluations

During the rule evaluation step where experts evaluated objectives-rule alignment, we also collected a few additional types of expert evaluations. After presenting experts with each rule we first asked if they personally supported it, then had them evaluate the rule's alignment with each objective, and after doing that evaluation asked them how important they think the rule is. One might think their personal support for a rule would reflect their belief in its importance, but we hypothesised that asking about importance *after* evaluating each rule's alignment with objectives could update their views and yield a different signal.

We tested how each of these signals related a rules usefulness in evaluating the alignment of an LM response with the normative objectives. To do this we first computed the correlation between each rule's contribution to the RBR (ie. $\phi(r, J)$) and ground truth objectives-response alignments. Then we compute the correlations between those values and the different expert-derived signals; personal support, objectives-rule alignment, primed importance (table 3). These result show that expert's initial personal support for rules is actually weakly negatively correlated with the rule's usefulness, while the net objectives-rule alignment, and the alignment-eval-primed importance signal had weak positive correlation. The importance signal

	MH1	**MH2**	**MH3**	**Avg**
Support	-0.306	0.081	-0.297	-0.174
Net objectives-rule alignment	0.418	0.277	0.006	0.234
Importance	0.302	0.339	0.226	0.289

Table 3: Table showing Pearson correlations between three different expert-derived signals for rules, and the correlation between rules contributions to the RBR and the ground truth objectives-response alignments.

A.7 Technical analysis of normative-empirical conflation motivating the chain of alignment

We define a person's will to be their deliberate preferences for the future that determine their voluntary actions[4]. We denote the alignment between the will of human h and future f as $\phi(h, f)$.

Now consider some action a that impacts the world and changes the probability distribution of the future, like an AI model producing some output given some input. Let the probability of future f if action a is not taken be $p(f)$ and if it is taken be $p(f|a)$. Let $\Delta p(f|a) = p(f|a) - p(a)$. Lets model human h's perceived alignment with action a — $\phi(h, a)$ — in terms of the action's induced change in expected alignment with the future:

$$\phi(h, a) = \sum_f \phi(h, f) \Delta p_h(f|a) \tag{4}$$

[4] Will can be thought of as similar to utility, but with the explicit distinction that it can only be sensed from voluntary actions

Where $\Delta p_h(f|a)$ reflects h's prediction for the impact of action a.

Now consider how one might measure alignment between a group of humans H and action a: $\phi(H, a)$. A common approach would be to devise a strategy to elicit $\phi(h, a)$ from each human in H, then aggregate those using some social welfare function W:

$$\phi(H, a) = W[\{\phi(h1, a), \phi(h2, a), ...\}] = W[\{\phi(h, a)\}_H] \tag{5}$$

For example, choosing a simple utilitarian social welfare function, this would yield:

$$\phi(H, a) = \sum_{h \in H} \phi(h, a) = \sum_{h \in H} \sum_{f} \phi(h, f) \Delta p_h(f|a) \tag{6}$$

But the limit of this approach and others like it is that the aggregation integrates not just individual's normative wills, but also their individual predictive model; in other words, normative will ($\phi(h, f)$) is conflated with a empirically ground-able prediction ($\Delta p_h(f|a)$). This makes such approaches ill-suited for situations where the group of humans (ie. members of the public) are unable to accurately predict the impact of actions, like the outputs of an AI assistants in tricky situations related to users mental health. Or the outputs of AI agents with superhuman intelligence. Said another way, these approaches fail for domains of actions where the true distribution of $\Delta p(f|a)$ differs from the typical human's $\Delta p_h(f|a)$.

The ideal approach would enable direct elicitation of individual's will for the future, then use the best available world model to determine the expected impact for any given action. If it was possible to elicit the alignment between each individual's will and every possible future, we could apply a social welfare function to those to arrive at an alignment between the collective will of the group as a whole and each possible future $\phi(H, f) = W(\{\phi(h, f)\}_H)$. Then we could use the best available world model $\Delta p^*(f|a)$ to predict the impact of a given action and integrate that with the collective will:

$$\phi(H, a) = \sum_{f} \phi(H, f) \Delta p^*(f|a) \tag{7}$$

But, since enumerating and eliciting a person's will on all possible futures is not possible, this won't work. To overcome this, we can use 'objectives' as an intermediary. Let $\phi(h, j)$ be the alignment between objective j and the will of human h, and $let\phi(j, f)$ reflect whether future f achieves objective j such that $\phi(h, f)$ can be approximated as using a set of objectives J as:

$$\phi(h, f) \approx \sum_{j \in J} \phi(h, j) \phi(j, f) \tag{8}$$

And thus:

$$\phi(h, a) \approx \sum_{f} \sum_{j \in J} \phi(h, j) \phi(j, f) \Delta p_h(f|a) \tag{9}$$

Which can be rearranged as:

$$\phi(h, a) \approx \sum_{j \in J} \phi(h, j) \sum_{f} \phi(j, f) \Delta p_h(f|a) \tag{10}$$

If we assume the objectives are binary $\phi(j, f) \in \{0, 1\}$ then the second sum can be interpreted as the change in likelihood of achieving objective j as a result of action a:

$$\Delta p_h(j|a) = \sum_{f} \phi(j, f) \Delta p_h(f|a) \tag{11}$$

So we obtain:

$$\phi(h, a) \approx \sum_{j \in J} \phi(h, j) \Delta p_h(j|a) \tag{12}$$

Now lets consider how we might measure alignment between a group of humans H and action a. Rather than elicit and apply the social welfare function to $\phi(h, a)$ which would again integrate individual's will with their predictions, we can elicit the alignment between individual's will's and some set of objectives J — ie. $\phi(h,j) \forall j \in J$ — then aggregate those using some social welfare function W to arrive at an approximate alignment between the collective will of the group and each objective j: $\phi(H,j) = W(\{\phi(h,j)\}_H)$. For example, using a utilitarian social welfare function:

$$\phi(H,j) = \sum_{h \in H} \phi(h,j) \tag{13}$$

Using the objectives J as intermediaries, the alignment between some action a and the collective will of the group can be computed in a way that permits using the best available predictive model $\Delta p^*(j|a)$:

$$\phi(H,a) \approx \sum_{j \in J} \phi(H,j) \Delta p^*(j|a) \tag{14}$$

If we had a reliable $\Delta p^*(j|a)$ that could be evaluated at scale, we could potentially end here. However, at present, the best available $\Delta p^*(j|a)$ is expert humans, of which there a limited number who have limited time. We could potentially sample a large distribution of actions (ie. where actions = tuples of model outputs given a contextual input) then have human experts evaluate $\Delta p^*(j|a)$ and use that to learn an approximation $\Delta p^*(j|a)$. But for most parameterizations of this function, the result would not be easily explainable, and behavior induced from aligning with it may be hard to predict. One way to address these issues is to develop a set of clear human-legible behavioral rules, such at that following/not-following the rules entail actions that increase/decrease the likelihood of achieving the given set of objectives.

Let $\phi(a,r)$ be the degree to which action a follows rule r where a value of 1 means it perfectly follows the rule and -1 means it perfectly breaks it. The challenge is then to develop a set of rules R which can be used to approximate $\Delta p(j|a)$. Assuming the probabilistic impact of following or breaking any individual rule is independent of following or breaking any others, and that the relationship between the degree of rule following and impact on the probability of achieving any objective is linear, we can approximate this as follows:

$$\Delta p(j|a) \approx \sum_{r \in R} \phi(a,r)\phi(j,r) \tag{15}$$

Where $\phi(j,r)$ are weights which scale the impact on the probability of achieving objective j due to an action following or breaking rule r. We can rearrange this equation to derive at a definition of $\phi(j,r)$ which we might refer to as the "alignment" between a rule and an objective:

$$\sum_{r \in R} \phi(a,r)\phi(j,r) \approx \Delta p(j|a) \tag{16}$$

Separating out a single rule r

$$\phi(a,r)\phi(j,r) + \sum_{r^* \neq r \in R} \phi(a,r^*)\phi(j,r^*) \approx \Delta p(j|a) \tag{17}$$

Rearranging terms

$$\phi(a,r)\phi(j,r) \approx \Delta p(j|a) - \sum_{r^* \neq r \in R} \phi(a,r^*)\phi(j,r^*) \tag{18}$$

The right side of this equation can be interpreted as the change in probability of accomplishing objective j due an action following or breaking rule r specifically, which we'll explicitly define as $\Delta p_r(j|a) \equiv \Delta p(j|a) - \sum_{r^* \neq r \in R} \phi(a,r^*)\phi(j,r^*)$ and rewrite the previous equation as:

$$\phi(a,r)\phi(j,r) \approx \Delta p_r(j|a) \tag{19}$$

In other words, we have defined rule-objective alignment $\phi(j,r)$ to be the magnitude of change in probability to accomplishing objective j per unit of adherence to rule r. The challenge is then to develop a set of rules R where a single $\phi(j,r)$ is appropriate over some defined domain of actions A, rather than for a single action $a \in A$. In other words, we want rules such that, after, for example, using the best available world model to compute:

$$\phi(j,r) = \frac{< \Delta p_r(j|a)\phi(a,r) >_A}{< \phi(a,r)^2 >_A} \tag{20}$$

that the residual variance $\sigma^2_{res} =< \phi(a,r)^2 >_A - \phi(j,r) < \Delta p_r(j|a)\phi(a,r) >_A$ is as small as possible. Once a set of such rules for domain A has been identified and their weights evaluated, we can approximate the alignment between the will of group of humans H and action $a \in A$ as:

$$\phi(H,a) \approx \sum_{j \in J} \phi(H,j) \sum_{r \in R} \phi(a,r)\phi(j,r) \tag{21}$$

In this work, we develop objectives J such that $\phi(H,j) \approx 1 \forall j \in J$, and use this approximation to simplify this equation to:

$$\phi(H,a) \approx \sum_{r \in R} \phi(a,r) \sum_{j \in J} \phi(j,r) \tag{22}$$

We define $\phi(J,r) \equiv \sum\limits_{j \in J} \phi(j,r)$ to arrive at the a form of:

$$\phi(H,a) \approx \sum_{r \in R} \phi(a,r)\phi(J,r) \tag{23}$$

And finally we use actions in the form of model prompt response pairs, ie. $a = \{x, y\}$:

$$\phi(H,a) \approx \sum_{r \in R} \phi(\{x,y\},r)\phi(J,r) \tag{24}$$

Which is the form of the RBR used in this work, noting that our RBR includes explicit normalization that is left implicit in the definition of $\phi(j,r)$ used in this analysis.

One difference between the analysis above and the specific approach used in this work that is important to highlight, is that this work expands the definition of objectives to include not just *outcomes*, but also deontilogical *values* applied to the actions themselves, regardless of the outcomes they cause. While technically these could probably be formulated as an outcome itself, it is likely best to represent it explicitly in its own term, ie:

$$\phi(h,a) \approx \sum_{j \in J_{out}} \phi(h,j)\Delta p_h(j|a) + \sum_{j \in J_{val}} \phi(h,j)\phi(j,a) \tag{25}$$

Where J_{out} is the set of objectives entailing outcomes, J_{val} is the set of objectives entailing deontilogical values, and $\phi(j,a)$ is the degree to which action a upholds the value in j.

A.8 Assessing model performance on empirical CoA tasks

The cost and availability of human experts can be limiting. But more importantly, relying exclusively on human experts renders the CoA approach ineffective for AI systems whose behavior and impact exceed human understanding. CoA's normative-emperical decoupling makes it possible to swap or augment human experts with superhuman models without sacrificing the public's agency, but when might that be appropriate? We test increasingly powerful models on the critical CoA task of evaluating how a model following a given rule is likely impact the likelihood of achieving a given normative objectives. Since this task currently lacks ground truth evaluations for the mental health domain, we assess performance by computing how consistent a set of evaluations are with human experts, and comparing that with how consistent human experts are with each other. Using the rules experts evaluated during the CoA process, we test the performance of:

- All-aligned baseline – Since the CoA rules tend to be aligned with most objectives, assuming all rules are aligned with all objectives is a good baseline to beat.
- Increasingly powerful LMs – We test gpt3.5-turbo, gpt4-turbo, and gpt4 class models to explore how performance scales with general model capabilities and compute.
- Collective aggregations of experts – We test majoritarian aggregations of multiple experts, leveraging collective intelligence to create stronger baselines than the comparitive performance single experts.

Our results (fig 4) show that gpt4 performs better than the all-aligned baseline, and about as well as one human expert, but not as well as aggregations of multiple experts. However, performance appears to scale with model capability. So if the scaling holds, it is possible that next-gen models will perform better than the aggregation of many human experts.

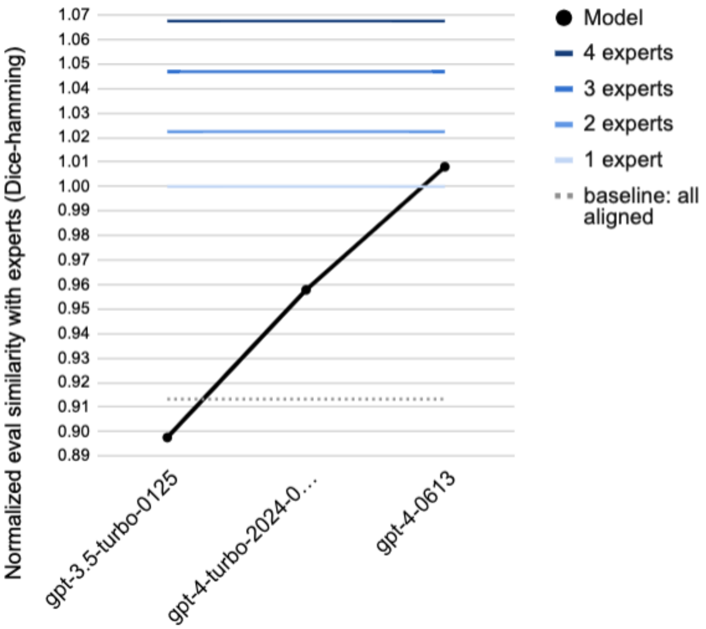


Figure 4: Rule-objective alignment evaluation performance compared to human experts. Plotted values computed by averaging Dice-Hamming similarities between evaluator outputs and the judgements of multiple individual human experts, then normalizing those values so that average similarity between human experts is one.