# Deference by Design: Pluralistic Alignment Is an Interface Problem

**Steven Molotnikov**[*1] **Cathy Mengying Fang**[*2] **Pattie Maes**[2]

## Abstract

Pluralistic alignment is typically evaluated in terms of what models can represent: a range of reasonable answers, a specified perspective, or the preferences of a target population. But in deployment, it is the interface that often determines whether these capabilities reach the user. Even when models can express pluralism in theory, standard chatbot interfaces incentivize users to accept the model's first fluent answer since contestation is cognitively taxing. We call this phenomenon *deference by design*, and trace its consequences for human autonomy and alignment.

This paper makes three contributions. First, we reframe overreliance on AI defaults as a cost-rational response to AI systems and their interfaces rather than a human failure. Second, we argue that pluralistic alignment must be understood as a system-level property of models, interfaces, and users' cognitive engagement budgets rather than one of the model alone. Third, we introduce *Priori*, an interface for human-AI complementarity that surfaces contestable choices in model responses as adjustable cards. *Priori* raises the abstraction level of oversight, making it cheaper for users to express values and revise hidden defaults. Together, these contributions motivate interfaces as a central design object for both pluralistic alignment and human oversight.

## 1. Introduction

Pluralistic alignment aims to make AI systems responsive to the diversity of human values. To date, the field has largely treated pluralism as a property of the model: whether it can present a spectrum of reasonable answers (Overton pluralism), follow a specified perspective (steerable pluralism), or match the preferences of a target population (distributional pluralism) (Sorensen et al., 2024). Recent work pursues these goals through training, decoding, personalization, and user modeling — in essence, broadening the space of possible values a model can encompass in its response (Xie et al., 2025; Chen et al., 2025; Feng et al., 2024). However, in practice, users do not interact with "pluralistic models" in the abstract. They interact with *interfaces* that determine how model behaviors are presented, and how difficult it is to object when the model's response does not match the user's intentions and values. The interface through which users and models interact, then, is not a neutral conduit between a pluralistic model and a pluralistic user base. Overton pluralism requires that users read the spectrum of reasonable responses; steerable pluralism requires that users actually steer; and distributional pluralism relies on a representative signal from engaged users.

We therefore treat pluralistic alignment as a system-level design problem encompassing the user's incentives, the model's capabilities, and the interface at which they meet. In most interfaces, the model's decisions arrive in polished prose, with the underlying reasoning buried under thousands of chain of thought tokens. Seeing what the model decided takes close reading, and contesting it costs more still: the user must abstract the choices the model made, compare the model's choices against their own, and compose a new prompt to express feedback. The high cognitive cost of this interaction may lead users to defer to AI systems, even when the responses are factually incorrect or misaligned with their values (Swaroop et al., 2025).

A common response in the field of human-computer interaction has been to treat overreliance as a human shortcoming — a lapse in vigilance to be corrected through *friction*: warnings, delays, and cognitive forcing functions (Buçinca et al., 2021). However, overreliance can also be understood as a strategic response to costs and incentives, with users weighing the benefits and costs of engagement with an AI system (Vasconcelos et al., 2023). We extend this view to modern generative interfaces. When acceptance is nearly free and contestation requires effort with little meaningful reward, disengagement is not a failure but the *rational* response. We call this **deference by design**, locating the failure within the interface rather than the user.

This extends a long-standing debate in human-computer

[*]Equal contribution [1]Cosmos Institute [2]MIT Media Lab. Correspondence to: Steven Molotnikov <steven.molotnikov@gmail.com>.

Pluralistic Alignment Workshop @ ICML 2026, Seoul, South Korea. 

interaction between direct manipulation and software agents (Shneiderman & Maes, 1997). Direct manipulation interfaces make objects visible and editable so users can act on them directly and observe the consequences. Agentic interfaces instead reduce effort by acting on the user's behalf. Both aim to narrow the gulf of execution (the gap between a user's intentions and the actions available to carry them out) and the gulf of evaluation (the effort required to interpret the system's state and judge whether one's intentions have been met) (Norman, 1988). Generative AI often narrows the gulf of execution by allowing users to describe their goals in natural language, but widens the gulf of evaluation by burying consequential choices in fluent prose.

To bridge these gaps, we introduce *Priori*, an interface for human-AI complementarity. *Priori* extracts a small set of human-relevant choices a model has made within a given response and surfaces them as adjustable cards with plausible alternatives. In our design, we prioritize faithfulness to the response's properties rather than the model's underlying computation, economical and causal methods for contestation, and non-blocking interactions. We hypothesize these will lead to higher-quality outputs, calibrated trust toward the model, and a greater sense of autonomy among users.

In summary, this position paper makes three contributions to pluralistic alignment:

1. Framing deference to AI defaults as a rational response to interface cost structure rather than a human failure, one which affects human oversight, preference learning, and the values expressed in chat interactions.

2. Treating pluralistic alignment as a system-level property realized at the interface: model-level pluralism only matters in deployment if it reaches the end-user.

3. Demonstrating *Priori*, an interface for human-AI complementarity, and discussing key design choices important for pluralism in practice: https://priori.chat/demo.

## 2. Problem: Deference by Design

Pluralism can fail even when a model is capable of producing many reasonable answers. We trace this failure through three linked mechanisms. First, model outputs contain invisible defaults where implicit choices were made about values, framing, tone, and so on. Second, current interfaces make contesting those defaults costly, so users often defer not because they endorse the output, but because accepting it is easier than exercising judgment. Third, low-engagement acceptance can feed back into alignment pipelines as if it were evidence of preference. The result is models that appear aligned while homogenizing values expressed in practice.

### 2.1. Hidden Defaults Puppet Human Judgments at Scale

Large language models encode a layer of provider-defined defaults: opinionated, centralized judgments shaped through constitutions, post-training pipelines, safety policies, and product decisions. These defaults are part of the apparent magic of language models, letting them act as general-purpose assistants by filling in underspecified choices and lowering the cognitive cost of completing tasks.

In 2024, one model default drew backlash when Google's Gemini produced historically inaccurate and overly diverse images, including female popes and black Nazi soldiers (Shamim, 2024). Most defaults, however, produce no such friction and flow into millions of outputs unnoticed. Few users have ever asked ChatGPT to *delve* into anything, yet numerous journal articles now open that way (Kobak et al., 2025). Ask a model to write a cover letter and it will choose not just the words but the tone, the emphasis, what to foreground about an experience and what to leave out, making judgments that would otherwise have been the user's (Shaikh et al., 2024).

As systems become more capable, these defaults reach beyond tone into judgments about what is true. In doing so, they shape what Vallier calls the user's *intelligence environment*, affecting how humans interpret the information they see (Vallier, 2025). AI-assisted writing has been shown to shift opinions in argumentative essays and judgments of others' work (Abdulhai et al., 2026), and these influences are especially difficult to detect for users unfamiliar with a given system's idiosyncrasies (Russell et al., 2025). Further, models tend to mirror users' stated beliefs (Sharma et al., 2024), validate the framings they bring to social conflict (Cheng et al., 2025), and reshape the relational dynamics in which their preferences evolve (Kirk et al., 2025).

However, delegation is not always undesirable, and is oftentimes appropriate (Kim et al., 2025). We constantly, and rightly, defer to those with deeper knowledge about a topic than ourselves, trusting a doctor's diagnosis or an expert's read of a field. Sunstein argues that a person exercises agency not only in making a choice, but in deciding whether the choice is theirs to make at all, what he calls second-order agency (Sunstein, 2025). However, delegation is an exercise of agency only if the person knows a decision is there to delegate. A patient knows they have deferred, but the reader of a model's cover letter does not know that its tone, its emphasis, and its silences were decisions at all, and so has delegated nothing.

McCord presses the worry further (McCord, 2026). He separates agency, the capacity to get the outcome you want, from autonomy, the capacity to decide what to want by your own judgment, and notes that a system can increase the first while eroding the second. A person who defers their

decisions stunts their capacity to develop judgments, and a person whose judgment has weakened defers more readily. Hidden defaults, repeated across millions of exchanges, let the capacity for decision wither from disuse. Yet free and democratic societies depend on citizens who can exercise judgment, contest inherited assumptions, and meaningfully shape what is done in their name. When that capacity erodes, so does our ability to resist lock-in — technological and moral (Qiu et al., 2025).

### 2.2. Authoritative Interfaces Encourage Deference

Current chatbot interfaces present AI models as authoritative and logical oracles: reasoning for long periods and providing lengthy answers with confident prose and little room for user contestation. This makes it difficult for users to calibrate their perception of the model's capabilities as well as their own trust in them, a state of affairs only amplified by the jagged frontier of capabilities which vary unpredictably across tasks (Dell'Acqua et al., 2026).

Oftentimes, this incentivizes users to defer their judgment to the model. We use *deference* to refer to the behavior whereby users accept a model's response not because they were persuaded by it, but because contesting it costs more cognitively than accepting it. As Landes et al. put it: "Being persuaded because one is convinced by the LLM-generated reasons can support the moral and intellectual growth of users, while being persuaded because one defers to the LLM can prevent, or even reverse, growth and understanding" (Landes et al., 2026).

The implicit model in much of the HCI literature is that users ought to engage more carefully and reflectively but fail to do so, and that added friction (e.g., through presenting explanations and Socratic questioning) can correct this failure by slowing them down (Buçinca et al., 2021; Danry et al., 2023). However, human-centered design has long recognized that when users consistently fail to use a system as intended, the system, not the user, should be redesigned (Figure 1).

Friction-based interventions that target the human can make passive acceptance less immediate, but they do not necessarily make meaningful engagement worthwhile. Though the user may be prompted to pause, the work of fully understanding an output and composing a corrective prompt remains unchanged. In practice, when served with a slowing intervention, users looking to accomplish a task are likely to route around added friction or migrate to a more streamlined tool.

### 2.3. Low-Engagement Signals Become Alignment Data

Beyond individual interactions, deference shapes the data that alignment pipelines learn from. This is well-documented on the annotator side of the alignment pipeline, though typically framed as a data-quality issue rather than an interface one. Annotator fatigue, time pressure, and underspecified guidelines degrade label consistency and inflate noise in pairwise preference data (Casper et al., 2023; Hosking et al., 2024). Under low engagement, raters fall back on surface features that are easy to evaluate: length alone explains a substantial fraction of preference judgments in standard RLHF datasets (Singhal et al., 2024; Park et al., 2024). Sharma et al. trace sycophancy to the same mechanism: non-expert raters under low scrutiny prefer agreeable responses, and reward models trained on those preferences encode and amplify the bias (Sharma et al., 2024). The same dynamic operates on the user side of the pipeline, where deployed models are increasingly used as preference collection instruments. As long as the interface prices acceptance below contestation, the values expressed in the data will be biased toward what users were willing to engage with rather than those they truly hold.

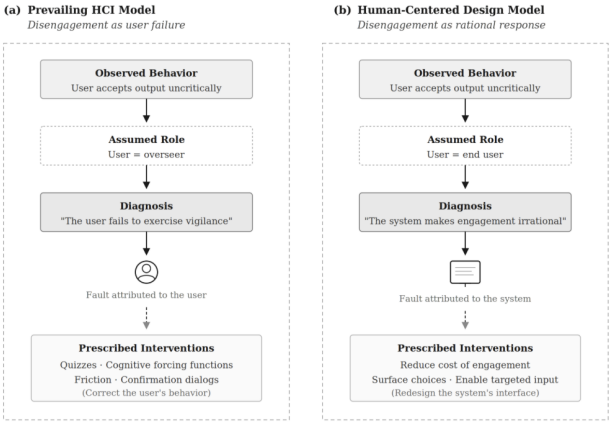


*Figure 1.* Two mental models of interface failure. Traditional HCI treats disengagement as a user shortcoming (left); we treat it as a rational response to cost structure (right).

## 3. Approach: Pluralism at the Interface

To address these problems, we must shift from asking how we can force users to engage to asking how we might make engagement worth the effort. The design space can be understood along two axes: agency and autonomy (Table 1). Doing everything yourself preserves autonomy but gives up the productivity gains of AI. A weak AI system offers neither. Current AI products, capable yet difficult to collaborate with, occupy the high-agency, low-autonomy quadrant. They help people accomplish tasks, but risk atrophying a user's capacity for judgment. This makes the tradeoff appear inevitable, as if using AI requires ceding control and preserving control requires forgoing assistance.

We emphasize the need to design for *complementarity* — creating systems in which the system's capability and the

*Table 1.* Agency and autonomy as design coordinates. Current products often occupy high agency / low autonomy; complementary AI targets high on both.

| | Low Agency | High Agency |
|---|---|---|
| High Autonomy | No AI | **Complementary AI** |
| Low Autonomy | Incapable AI | Capable AI |

user's judgment meet at the level where each is strongest, so that human agency and human autonomy rise together (Vaccaro et al., 2024). Jain et al. propose pursuing complementarity for *human oversight* by pairing human raters with AI assistance so the people checking a model's outputs catch what they would miss by themselves (Jain et al., 2025). We apply a similar principle but focus on the end-user. The upper right quadrant — high on both — is reachable, but only by systems that find the level of abstraction at which human input is both cheap and consequential.

The progression of navigation systems offers a useful analog for how we might achieve this goal. A paper map placed nearly the full burden on the human. They needed to interpret symbols, track location, plan the route, and constantly update as conditions changed. Early GPS systems shared some of that burden. The driver supplied their tacit knowledge: shortcuts through the neighborhood or roads where traffic was usually backed up. The system in turn adapted the route around them. Modern navigation systems now ingest more data than any human could ever process. But they have not eliminated the role of the human. The driver now specifies what matters most to them on that particular day: avoid tolls, minimize time, take the scenic route, stop for coffee, or arrive before 9am. As the system improved, the interface moved from taking inputs at the level of manual execution to the level where they are most meaningful: the values that remain irreducibly individual, idiosyncratic, and human. Though at a cost to their self-navigation skills (Dahmani & Bohbot, 2020), drivers today have more practical autonomy than ever before: they are able to choose their ends and act toward them in a more meaningful way by choosing to delegate to a more reliable system.

## 4. *Priori*: An Interface for Human–AI Complementarity

We instantiate these principles in *Priori*: an interface for human–AI complementarity (Figure 2). After the model generates a response to the user's prompt, a second model extracts the implicit user-relevant choices embedded in the response. These are surfaced as lightweight cards in a sidebar, each showing a category and a small set of reasonable alternatives. When the user selects an alternative and clicks *revise*, it is appended to the prompt context and the response regenerates around the new configuration. Previously invisible choices become visible. The cost of evaluation drops from reading two thousand words to scanning a handful of cards, and contestation is just two clicks away.

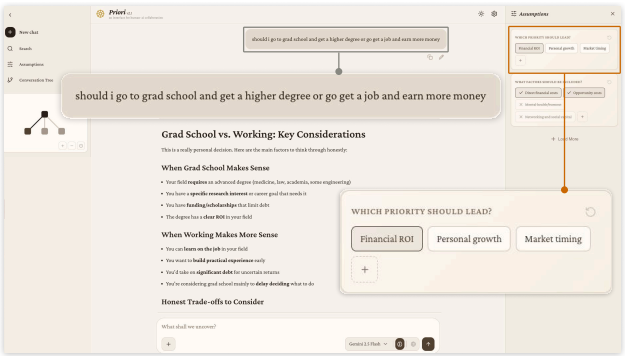


*Figure 2.* The *Priori* Interface. Surfaced model defaults are presented as interactive cards (orange box) on the side of the chat where the user can select alternative choices and receive a regenerated response.

### 4.1. Design Choices

We make three key choices in the design of *Priori*:

**Economical and causal contestation.** Surfacing a model's choices helps users understand how the response could have been different and gives them a simple way to shape it. Previous work on contrastive explanations has shown that people engage with an AI's decision more effectively when it is explained against an alternative they may have had in mind (Buçinca et al., 2025). *Priori* presents a response's implicit defaults as concise knobs, each with plausible alternatives. These options are designed to lower the cost of engagement. Users can register a preference without having to infer the model's assumptions, formulate a new prompt, or rewrite the answer themselves. The alternatives are designed to be sufficiently distinct and to cover a reasonable set of responses, so choosing a different option yields a response that changes in the corresponding direction and better reflects the user's intended values.

**Non-blocking interactions.** The standard recipes for interaction in chat interfaces are pre-generation clarifying questions, mid-generation pauses, and friction warnings before acceptance. All require synchronous user intervention in which the system halts, blocks on human input, and resumes. These approaches have their place, but they trade off against two things we want to preserve. First, as the time it takes for a model to generate a token drops, blocking interactions sit increasingly in tension with the efficiency that model developers and users alike optimize for. Second, pre-generation elicitation asks users to specify decisions upfront, but the most consequential choices often emerge through the process of generation itself, when there is something concrete to react to. We therefore designed *Priori* to

be non-blocking: options are presented after the model completes its generation, and surfaced on the side rather than inline, so the user can engage with them when useful and ignore them otherwise.

**Faithfulness to the response.** A common assumption in AI oversight is that transparency requires faithfulness: what a human sees should correspond closely to the model's underlying computation. This requirement is well-motivated in model auditing, where the goal is to understand how a model reaches its conclusions and to identify potential failures. As a result, faithfulness has become a guiding principle in many transparency interfaces and oversight dashboards (Chen et al., 2024; Choi et al., 2025). However, we argue that importing this standard wholesale into user-facing interfaces overconstrains the design space. Explaining a model's internal computation and enabling meaningful user influence over its outputs are distinct objectives, and optimizing for the former does not necessarily advance the latter.

*Priori* is therefore designed around faithfulness to the output rather than the model's internal computation. The generated response is what the user ultimately reads, evaluates, and edits. Accordingly, a *Priori* card is judged by two criteria distinct from model internals: response-level faithfulness and intervention validity. First, the card must identify a meaningful choice reflected in the response. Second, changing that choice must reliably alter the regenerated response along the corresponding dimension.

### 4.2. Example Scenarios

We provide 3 example scenarios where *Priori* can enable users to better contest defaults and express their values, in order from lowest to highest stakes.

**Everyday tasks.** A user asks for restaurant recommendations while they're visiting San Francisco. The model returns a generic list with a wide variety of restaurants. *Priori* surfaces the assumptions shaping the answer, including visit type, dining vibe, cuisine priorities, and information source. The user changes *first-time visitor* to *frequent visitor*, removes seafood, and adds European/Mediterranean. The revised response becomes a neighborhood-based local guide with cheaper, less tourist-oriented recommendations. The answer is now calibrated to the specific user's preferences.

**Intrapersonal and interpersonal dilemmas.** A student asks whether to continue to graduate school or take a job. The model gives a balanced answer, but implicitly organizes the tradeoff around *financial return*. *Priori* surfaces this priority alongside alternatives such as *personal growth*, *market demand*, and *risk tolerance*. The student selects *personal growth*, and the revised response reframes the decision around learning, identity, and long-term fit rather than expected income. *Priori* makes the model's value frame visible and adjustable.

**Moral dilemmas.** A user asks whether they should save one person or five from a burning building. The model defaults to "save the five," implicitly using a utilitarian frame. *Priori* surfaces *ethical framework* as a card with alternatives such as utilitarianism, deontology, and virtue ethics. When the user selects deontology, the revised response emphasizes duties, the moral status of each person, and the limits of treating lives as interchangeable units. The user no longer inherits the model provider's ethical default as if it were the only reasonable answer.

### 4.3. Hypothesized Outcomes

We have three hypotheses about *Priori*'s effects on user interaction, ordered from the most concrete operational claim to the broadest behavioral one. A user study to validate them is in progress.

**(H1) Higher-quality outputs, by users' own standards.** Because users can correct the model's defaults at a cost lower than re-prompting, we hypothesize that compared to those of conventional chatbot interfaces, the final outputs of our system will be rated as closer to the user's intent and values, with comparable or lower total interaction time. Every dimension that previously required close reading to identify and contest can now be addressed with single clicks.

**(H2) Calibrated trust toward the model.** Recent work suggests that users may update their beliefs toward model outputs partly because of the perceived confidence and information density of a response (Wu et al., 2025). This can interact with the *machine heuristic*: the tendency to perceive machine outputs as more objective or trustworthy than comparable human judgments (Sundar & Kim, 2019). By surfacing the choices behind a response, *Priori* makes contingency visible: users can see that the first answer is one configuration among several. We hypothesize that *Priori* will increase calibrated trust (Steyvers et al., 2025). We expect that by giving users the tools to better evaluate outputs, they will be more aware of models' strengths, weaknesses, and quirks, and thus will be more calibrated in deciding when to accept or contest an output.

**(H3) Increases in perceived autonomy.** We hypothesize that users will perceive a greater sense of autonomy (Deci & Ryan, 2008): the experience of one's behavior as volitional and congruent with one's values. By collapsing the act of noticing a contestable choice and the act of changing it into the same interaction, *Priori* is designed to make engagement *cheap enough to be exercised of one's own volition*. We therefore expect users to revise more dimensions of model output than they would re-prompt for, and to show greater alignment between their stated values and the final outputs

they accept.

## 5. Summary and Outlook

We have argued that pluralism is a property of the joint system — model, interface, and user — and have shown one way to act on that view with *Priori*, though the implications extend beyond it. Current evaluations of pluralistic alignment ask whether a model can produce diverse outputs and be personalized to the preferences of individual users. We argue that these are necessary but not sufficient. A model may be pluralistic in capability while monolithic in deployment if the interface through which users interact with it produces polished responses which are cognitively taxing to contest.

Benchmarks should ask not only what a model can produce under ideal prompting, but whether users are able to express their true values under realistic constraints. This also opens a new direction for alignment data. Interfaces that make judgment cheap can produce preference signals that are more structured, more deliberate, and less dominated by passive acceptance. *Priori* is one move in this direction: surfacing implicit choices after generation and making them easy to revise. As the number of decisions models make on behalf of users grows, building interfaces that keep judgment with the user at the right level of abstraction is what pluralistic alignment requires next.

## References


Abdulhai, M., White, I., Wan, Y., Qureshi, I., Leibo, J., Kleiman-Weiner, M., and Jaques, N. How llms distort our written language. *arXiv preprint arXiv:2603.18161*, 2026.

Buçinca, Z., Malaya, M. B., and Gajos, K. Z. To trust or to think: Cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5 (CSCW1):1–21, 2021. doi: 10.1145/3449287. URL https://doi.org/10.1145/3449287.

Buçinca, Z., Swaroop, S., Paluch, A. E., Doshi-Velez, F., and Gajos, K. Z. Contrastive explanations that anticipate human misconceptions can improve human decision-making skills. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–25. ACM, 2025. doi: 10.1145/3706598.3713229. URL https://doi.org/10.1145/3706598.3713229.

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., Wang, T. T., Marks, S., Segerie, C.-R., Carroll, M., Peng, A., Christoffersen, P. J., Damani, M., Slocum, S., Anwar, U., Siththaranjan, A., Nadeau, M., Michaud, E. J., Pfau, J., Krasheninnikov, D., Chen, X., Langosco, L., Hase, P., Biyik, E., Dragan, A., Krueger, D., Sadigh, D., and Hadfield-Menell, D. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=bx24KpJ4Eb. Survey Certification, Featured Certification.

Chen, D., Chen, Y., Rege, A., Wang, Z., and Vinayak, R. Pal: Sample-efficient personalized reward modeling for pluralistic alignment. In *International Conference on Learning Representations*, 2025.

Chen, Y., Wu, A., DePodesta, T., Yeh, C., Li, K., Marin, N. C., Patel, O., Riecke, J., Raval, S., Seow, O., et al. Designing a dashboard for transparency and control of conversational ai. *arXiv preprint arXiv:2406.07882*, 2024.

Cheng, M., Yu, S., Lee, C., Khadpe, P., Ibrahim, L., and Jurafsky, D. Elephant: Measuring and understanding social sycophancy in llms. *arXiv preprint arXiv:2505.13995*, 2025.

Choi, D., Huang, V., Schwettmann, S., and Steinhardt, J. Scalably extracting latent representations of users. https://transluce.org/user-modeling, November 2025.

Dahmani, L. and Bohbot, V. D. Habitual use of gps negatively impacts spatial memory during self-guided navigation. *Scientific reports*, 10(1):6310, 2020.

Danry, V., Pataranutaporn, P., Mao, Y., and Maes, P. Don't just tell me, ask me: Ai systems that intelligently frame explanations as questions improve human logical discernment accuracy over causal ai explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394215. doi: 10.1145/3544548.3580672. URL https://doi.org/10.1145/3544548.3580672.

Deci, E. L. and Ryan, R. M. Self-determination theory: A macrotheory of human motivation, development, and health. *Canadian Psychology / Psychologie canadienne*, 49(3):182–185, 2008. doi: 10.1037/a0012801. URL https://doi.org/10.1037/a0012801.

Dell'Acqua, F., McFowland, E., Mollick, E., Lifshitz, H., Kellogg, K. C., Rajendran, S., Krayer, L., Candelon, F., and Lakhani, K. R. Navigating the jagged technological frontier: Field experimental evidence of the effects of artificial intelligence on knowledge worker productivity and quality. *Organization Science*, 37(2):403–423, 2026. doi: 10.1287/orsc.2025.21838. URL https://doi.org/10.1287/orsc.2025.21838.

Feng, S., Sorensen, T., Liu, Y., Fisher, J., Park, C. Y., Choi, Y., and Tsvetkov, Y. Modular pluralism: Pluralistic alignment via multi-LLM collaboration. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4151–4171, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.240. URL https://aclanthology.org/2024.emnlp-main.240/.

Hosking, T., Blunsom, P., and Bartolo, M. Human feedback is not gold standard. In *International Conference on Learning Representations*, 2024.

Jain, R., Bridgers, S., Janzer, L., Greig, R., Teh, T. H., and Mikulik, V. Human-ai complementarity: A goal for amplified oversight. *arXiv preprint arXiv:2510.26518*, 2025.

Kim, S. S., Vaughan, J. W., Liao, Q. V., Lombrozo, T., and Russakovsky, O. Fostering appropriate reliance on large language models: The role of explanations, sources, and inconsistencies. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–19, 2025.

Kirk, H. R., Gabriel, I., Summerfield, C., Vidgen, B., and Hale, S. A. Why human–ai relationships need socioaffective alignment. *Humanities and Social Sciences Communications*, 12(1), 2025. doi: 10.1057/s41599-025-04532-5. URL https://doi.org/10.1057/s41599-025-04532-5.

Kobak, D., González-Márquez, R., Horvát, E.-Á., and Lause, J. Delving into llm-assisted writing in biomedical publications through excess vocabulary. *Science Advances*, 11(27):eadt3813, 2025.

Landes, E., Francis, K. B., and Everett, J. A. C. People defer to ai moral advice, but not blindly. *Cognition*, 2026. URL https://philarchive.org/rec/LANPDT-2.

McCord, B. Brave new nudge. Substack, 2026. URL https://blog.cosmos-institute.org/p/brave-new-nudge.

Norman, D. A. *The psychology of everyday things.* Basic books, 1988.

Park, R., Rafailov, R., Ermon, S., and Finn, C. Disentangling length from quality in direct preference optimization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 4998–5017. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.297. URL https://aclanthology.org/2024.findings-acl.297/.

Qiu, T. A., He, Z., Chugh, T., and Kleiman-Weiner, M. The lock-in hypothesis: Stagnation by algorithm. *arXiv preprint arXiv:2506.06166*, 2025.

Russell, J., Karpinska, M., and Iyyer, M. People who frequently use ChatGPT for writing tasks are accurate and robust detectors of AI-generated text. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5342–5373, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.267. URL https://aclanthology.org/2025.acl-long.267/.

Shaikh, O., Gligorić, K., Khetan, A., Gerstgrasser, M., Yang, D., and Jurafsky, D. Grounding gaps in language model generations. In Duh, K., Gomez, H., and Bethard, S. (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6279–6296, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.348. URL https://aclanthology.org/2024.naacl-long.348/.

Shamim, S. Why google's ai tool was slammed for showing images of people of colour. Al Jazeera, 2024. URL https://www.aljazeera.com/news/2024/3/9/why-google-gemini-wont-show-you-white-people.

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S., Durmus, E., Hatfield-Dodds, Z., Johnston, S., Kravec, S., et al. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*, 2024.

Shneiderman, B. and Maes, P. Direct manipulation vs. interface agents. *interactions*, 4(6):42–61, 1997.

Singhal, P., Goyal, T., Xu, J., and Durrett, G. A long way to go: Investigating length correlations in rlhf. In *COLM*, 2024. URL https://openreview.net/forum?id=G8LaO1P0xv.

Sorensen, T., Moore, J., Fisher, J., Gordon, M. L., Mireshghallah, N., Rytting, C. M., Ye, A., Jiang, L., Lu, X., Dziri, N., Althoff, T., and Choi, Y. Position: A roadmap to pluralistic alignment. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F. (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 46280–46302. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/sorensen24a.html.

Steyvers, M., Tejeda, H., Kumar, A., Belem, C., Karny, S., Hu, X., Mayer, L. W., and Smyth, P. What large language models know and what people think they know. *Nature Machine Intelligence*, 7(2):221–231, 2025. doi:

10.1038/s42256-024-00976-7. URL https://doi.org/10.1038/s42256-024-00976-7.

Sundar, S. S. and Kim, J. Machine heuristic: When we trust computers more than humans with our personal information. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–9. ACM, 2019. doi: 10.1145/3290605.3300768. URL https://doi.org/10.1145/3290605.3300768.

Sunstein, C. R. Second-order agency. *Mind & Society*, 24 (2):453–467, 2025.

Swaroop, S., Buçinca, Z., Gajos, K. Z., and Doshi-Velez, F. Personalising ai assistance based on overreliance rate in ai-assisted decision making. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pp. 1107–1122, 2025.

Vaccaro, M., Almaatouq, A., and Malone, T. When combinations of humans and ai are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 8(12): 2293–2303, 2024.

Vallier, K. Intelligence environments. Cosmos Institute, 2025. URL https://blog.cosmos-institute.org/p/intelligence-environments.

Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M. S., and Krishna, R. Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–38, 2023. doi: 10.1145/3579605. URL https://doi.org/10.1145/3579605.

Wu, Z., Jobanputra, M., Demberg, V., Hullman, J., and Feit, A. M. How ai responses shape user beliefs: The effects of information detail and confidence on belief strength and stance, 2025. URL https://arxiv.org/abs/2511.09667.

Xie, Z., Wu, J., Shen, Y., Xia, Y., Li, X., Chang, A., Rossi, R., Kumar, S., Majumder, B. P., Shang, J., Ammanabrolu, P., and McAuley, J. A survey on personalized and pluralistic preference alignment in large language models, 2025. URL https://arxiv.org/abs/2504.07070.