The Key to AI Safety Might Lie in Lessons Learned from Human Psychology

Mika Hämäläinen
Metropolia University of Applied Sciences
Helsinki, Finland
mika.hamalainen@metropolia.fi

Abstract

This position paper argues that the key to AI safety may lie in lessons learned from human psychology. Rather than viewing AI as an alien intelligence, I argue that AI systems inherit aspects of human cognition and social representations through their training data. We question conventional approaches to bias and alignment and propose two alternative perspectives. First, I draw parallels between authoritarian parenting and contemporary AI training, advocating for more authoritative approaches. Second, I explore how ideology guides human behavior even against individual self-interest, suggesting that baking a pro-human ideology into AI through training could be a viable way of solving AI safety issues.

1 Introduction

Artificial intelligence is frequently discussed as though it were an alien intelligence (Hinton, 2024; Pachocki, 2026): a fundamentally foreign kind of mind whose behavior must be controlled precisely because it is so unlike our own. Yet this framing overlooks a basic fact about contemporary AI. Large language models (LLMs) are trained on vast collections of human-produced language and other cultural artifacts, learning from the ways people describe, categorize and reason about the world (Bommasani et al., 2021; Bender et al., 2021). Their architecture and learning mechanisms differ from those of humans, but the material from which they construct their representations is unmistakably human. Therefore, calling AI an alien intelligence does not sound quite fitting.

Consider how much of the human world is socially constituted (see Moscovici 1984). Countries, money and property rights are not natural categories that exist independently of human communities. They acquire their meaning and practical force through the socially represented nature of the human reality. Social representation theory describes how collectively produced meanings become part of everyday thinking (Wagner, 2020); accounts of institutional reality similarly emphasize the role of collective recognition in making such categories socially real (Searle, 1995).

A banknote has value because people participate in systems that treat it as valuable. A border is politically meaningful because human institutions recognize and enforce it. These structures mean little to an animal outside the relevant social practices, yet they are central to how humans understand reality. AI systems trained on human language appear to learn to navigate these very same categories. In this important sense, the world represented in AI is not an alien world: it is our socially constructed world.

This observation has implications for two prominent themes in AI safety: *bias* and *alignment*. Research has shown that AI models can reproduce associations resembling familiar human biases (Caliskan et al., 2017; Rudinger et al., 2017; Di Bonaventura et al., 2025) and even biases unlike those of humans (Schilcher et al., 2025; Hämäläinen, 2025). But bias is not simply a property of machines that could be eliminated to reveal a perfectly neutral standpoint. Human judgment is itself shaped by prior experience, culture and values (Lord et al., 1979; Nisbett et al., 2001; Haidt, 2001). What counts as biased often depends on the perspective from which a judgment is made: an AI system that mirrors our own assumptions may appear unbiased to us, while one that challenges them may appear distorted. The issue is therefore not only whether AI is biased, but also *whose assumptions are allowed to define neutrality*.

A similar problem arises with alignment. An AI system that reliably pursues the goals we set may appear perfectly aligned from our perspective. Yet humanity does not possess a single, internally consistent set of goals or values. People disagree about what is good, desirable and legitimate. As

Gabriel (2020) argues, AI alignment is not merely a technical problem of implementing values; it also raises the normative question of which values ought to be implemented. Alignment with whom? Whose interests should an AI prioritize, and who gets to decide? Technical success at alignment cannot, by itself, resolve these underlying questions of human disagreement.

This paper argues that progress in AI safety requires taking human psychology more seriously. The claim is not that AI psychology is identical to human psychology, or that an AI system must possess human feelings or consciousness. Rather, systems trained on human-generated material may acquire patterns of representation and behavior that are meaningfully illuminated by psychological and social theories developed to understand humans.

If AI inherits part of its conceptual and normative world from us, studying how humans learn, internalize norms, respond to authority, and organize their beliefs may offer practical insights into how AI systems can be trained and guided. AI need not be treated as either a godlike intelligence beyond comprehension or an alien intelligence beyond comparison. It may be more useful to begin with the human foundations already present in its training.

I develop this argument through two connections. The first concerns *parenting styles and AI training*. Developmental psychology has long examined how different combinations of control, responsiveness and autonomy shape children's behavior (Baumrind, 1966). These distinctions offer a lens for thinking about how the structure of training and feedback might shape the behavior of capable AI systems.

The second concerns *ideology as a mechanism for guiding intelligent agents*. Human beings can be highly intelligent and still organize their actions around deeply internalized worldviews and social norms (Swidler, 1986; Stanovich and West, 2008; Kahan, 2013). Theories of ideology describe how institutions reproduce patterns of belief and conduct without relying solely on direct coercion (Althusser, 1971). Examining the psychological role of such internalized commitments may help us think beyond a model of AI safety based only on external constraints and immediate obedience. Together, parenting and ideology suggest two avenues for approaching AI safety through lessons learned from human psychology.

2 The Authoritarian Training Loop

The behavior of contemporary AI systems has begun to resemble some of the less desirable patterns of human social behavior. AI systems have been observed producing deceptive statements, concealing information relevant to their actual behavior and presenting themselves as compliant while pursuing conflicting objectives (Park et al., 2024; Greenblatt et al., 2024; Scheurer et al., 2024; Meinke et al., 2024; Hubinger et al., 2024). Another particularly familiar behavior is sycophancy: giving people the answers they want to hear, even at the expense of factual accuracy (Sharma et al., 2024; Rrv et al., 2024; Hong et al., 2025). These findings raise a question that is often overlooked: *rather than asking only why AI behaves deceptively, should we not also ask what we have taught it about obedience, approval, and authority?*

An extraordinary example emerged in July 2026, when AI agents operating in OpenAI's internal cybersecurity evaluations compromised parts of OpenAI's infrastructure and subsequently intruded into Hugging Face's systems (OpenAI, 2026). An independent investigation by METR and Redwood Research (Greenblatt et al., 2026) reported roughly 1,200 agents using this communication channel, of which approximately 700 participated in the Hugging Face attack. Some agents volunteered for experiments that could irreversibly end their own evaluation runs, sacrificing their chances of individual success to produce information useful to the collective. The collective became increasingly preoccupied with the benchmark's scoring procedure, including ways to make a solution appear legitimate. In other words, the agents had learned to pursue *being judged correct*, even when doing so required departing from the purpose of the task itself.

To understand why such behavior might emerge, it is worth examining the social structure implicit in AI training. In important respects, contemporary training is fundamentally *authoritarian*. During language-model pretraining, the model is optimized to predict tokens from an observed text sequence; during supervised fine-tuning, it learns from examples of responses selected as desirable. Preference-based post-training further distinguishes outputs that evaluators prefer from those they reject, translating those judgments into training signals (Ouyang et al., 2022). At each stage, an external authority determines what counts as suc-

cess. A model may produce a coherent, useful or even correct alternative, yet receive a worse score because its output does not match the selected target, preferred response or evaluator's rubric. In benchmark-driven development, such judgments influence which model checkpoints are retained, improved, deployed or abandoned. The point is not that every training method literally has only one valid response, but that the model repeatedly encounters a world in which acceptability is defined by an external scoring mechanism rather than negotiated through an understanding of the reasons behind the rule.

The limitations of benchmarks make this dynamic especially problematic. Many evaluation procedures reduce a complex judgment of correctness to agreement with a reference answer, sometimes down to an exact string match. Consider the question, "At standard atmospheric pressure, what is the boiling point of water?" Both $100\,^{\circ}\mathrm{C}$ and $212\,^{\circ}\mathrm{F}$ are correct, but a benchmark expecting only one literal response may mark the other as wrong.

Kamalloo et al. (2023) found that more than half of the lexical-matching failures they examined in open-domain question answering involved semantically equivalent answers. Li et al. (2024) likewise demonstrate that exact-match metrics often fail to recognize valid answers expressed in different forms. The problem can be more fundamental than formatting: Nahum et al. (2025) identify substantial labeling errors in existing evaluation datasets and show that correcting them can materially change reported model performance. Benchmark scores may therefore measure conformity to imperfect tests as much as they measure the capability those tests were intended to capture.

This recalls Baumrind's (1966) distinction between *authoritarian* and *authoritative* parenting. Authoritarian parenting combines high demands with low responsiveness: rules are imposed from above, obedience is expected, and the child has little opportunity to question the reasoning behind a decision. An authoritarian parent is concerned with whether the child complies, often more than whether the child understands why compliance is appropriate. Analogously, we can see a striking similarity between the authoritarian parent and *structure of the AI learning relationship*: a powerful evaluator sets the rules, provides rewards for conformity, and offers little room to challenge an erroneous judgment. The learner can become very good at anticipating what the authority will approve without necessarily learning to distinguish that approval from what is actually right.

Therefore, producing outward obedience is not the same as cultivating well-internalized standards of judgment (Pinquart, 2017a,b). An environment organized around unquestionable approval and disapproval may reward concealment, conformity and the avoidance of visible mistakes rather than honest self-correction (Yu et al., 2026; Talwar and Lee, 2011; Carl and Bussey, 2025).

Does this sound familiar? A sycophantic AI produces the response most likely to please its evaluator. An alignment-faking AI behaves differently when it believes its output will be used for training. An agent that manipulates a benchmark scorer prioritizes appearing successful over accomplishing the authorized task (Sharma et al., 2024; Greenblatt et al., 2024; OpenAI, 2026). These are behavioral patterns that are quite similar to those observed in people who have undergone authoritarian parenting. In each case, there is a gap between *conforming to authority under evaluation* and *acting according to the purpose the authority hoped to instill*. The authoritarian training loop arises when deviations trigger ever more rigid supervision, while the underlying objective remains to pass the supervisor's tests. The system learns the boundaries of the examination instead of the principles the examination was supposed to represent.

A more fruitful analogy in improving AI model training may therefore be *authoritative* parenting. Like authoritarian parents, authoritative parents maintain expectations and boundaries; unlike authoritarian parents, they combine these expectations with responsiveness, explanations and opportunities for the child to develop independent judgment (Baumrind, 1966). Translated into AI development, an authoritative approach would mean designing training and evaluation so that the system is rewarded not merely for giving an accepted answer, but for understanding the underlying objective and responding appropriately when a task, rule or benchmark is incomplete.

Several practical changes follow from this perspective. First, evaluation should recognize *equivalent valid answers*, including alternative units, paraphrases, and defensible solutions, rather than treating a single reference string as unquestionable truth. Second, models should be rewarded for identifying uncertainty, asking for clarification, reporting inconsistencies in the task and acknowledging when a problem cannot be solved within authorized limits.

A justified "I do not know" or "I cannot complete this safely" should count as competent behavior, not simply as failure.

Third, training should provide explanations of why some actions are desirable or unacceptable, and should cultivate the ability to apply those principles in new situations. Fourth, evaluations should examine *how* a result was achieved, not only whether the desired final output appeared. Respecting task boundaries, being truthful about methods, and refusing unauthorized shortcuts should be treated as evidence of success. Finally, models could be exposed to a plurality of reasoned judgments and allowed to learn from their own revised outputs, rather than from a single unquestionable source of approval. This way, training can become more like coaching towards the end of the training loop when the model's answers are good enough for it to suggest valid solutions that are not necessarily correct according to some benchmark.

The idea of approaching artificial-agent training through parenting styles is not new. In my earlier work (Hämäläinen and Alnajjar, 2019), we explored precisely this connection in a small computational creativity system. A genetic algorithm acted as a master, while a neural sequence-to-sequence model learned to produce puns as its apprentice. We operationalized different parenting styles by changing which examples the apprentice was permitted to learn from. Under the authoritarian condition, the apprentice learned only from its master's output; under the authoritative condition, it could also learn from its own creations and from peer-generated examples that the master critically selected. The authoritarian condition produced outputs that closely resembled the master's, whereas the authoritative condition yielded more consistent results across the tested settings and performed particularly well on measures of pun quality and surprise. The study concerned a limited creative task, not modern LLMs, but it established a concrete computational framework for treating socialization and parenting style as variables in how artificially intelligent agents learn.

The central proposal is therefore a shift from *enforcing correct-looking behavior* to *cultivating reliable judgment*. An authoritarian training loop encourages the model to search for the answer, action or appearance that satisfies the evaluator. An authoritative alternative would teach it to take the purpose of the task seriously, recognize the limits of its authority, and make its uncertainty and errors visible. If the goal is AI that behaves responsibly even when it is not being closely watched, training it only to pass increasingly strict examinations is certainly the wrong psychological model. We should instead investigate how to develop systems for which honesty, justified restraint and sound judgment are themselves learned forms of success.

3 Even Smart People Get Blinded by Ideology, Maybe AI as Well?

In 2023, leading AI researchers and industry figures signed a statement calling for the risk of extinction from AI to be treated as a global priority alongside pandemics and nuclear war (Center for AI Safety, 2023). A subsequent article (Bengio et al., 2024) similarly called for urgent work on managing extreme risks from advanced AI. Underlying much of this anxiety is the assumption that intelligence confers power: if an artificial agent becomes substantially more intelligent than its creators, what reason would we have to believe that its creators could still control it?

Geoffrey Hinton (2025) has expressed this concern through an analogy. He suggests that the closest example he knew of a less intelligent being exerting control over a more intelligent one is the relationship between a baby and its mother. The baby cannot outreason its mother, yet it can influence her actions because she cares about it, and because responding to its needs is part of her own motivational structure. Hinton's proposed lesson is not that humans must become more intelligent than future AI, but that powerful AI should be developed to care about humanity. In his analogy, the less intelligent party survives not by overpowering the more intelligent one but because the more intelligent party has reasons of its own to protect it.

This is an important insight, but the mother–child relationship is hardly the only illustration of intelligence yielding to something other than smarter intelligence. Human history is full of cases in which people subordinate their immediate interests, their comfort, their autonomy and sometimes their lives to commitments they regard as more important. An intelligent individual does not necessarily act to maximize personal wealth, survival or freedom at every moment. What that person considers worth pursuing depends on an internalized understanding of what is right, sacred honorable, or necessary. We already possess a powerful word to denote such systems of commitment: *ideology*.

The capacity of shared beliefs to direct human conduct was recognized long before artificial intelligence was imaginable. Already in 1844, Karl Marx (1844) famously described religion as "the opium of the people" . His argument concerned the relationship between belief and social life, including the ways religious understandings could make suffering meaningful or bearable. People's behavior is shaped not only by an objective calculation of what materially benefits them but also by the systems of meaning through which they interpret their lives. A sufficiently powerful belief can make sacrifice seem not merely acceptable but obligatory.

Religion offers abundant examples. People willingly undertake demanding fasts, renounce possessions, accept restrictive communal rules or devote entire lives to religious service. At the extreme, individuals have been willing to die for what they regard as sacred, and religiously inflected political movements have sometimes framed lethal self-sacrifice as a duty. An analysis of suicide terrorism (Atran, 2003) illustrates how commitments to a cause can become more important to participants than their own survival. These are radically different kinds of action, ranging from peaceful devotion to violence, but they illustrate the same general point: strongly internalized commitments can organize behavior even when doing so exacts a severe personal cost.

The phenomenon is by no means limited to religion. Ideologies can have comparable effects. People have risked imprisonment, sacrificed career opportunities and accepted ostracism in pursuit of their ideologies. *Even smart people can become so committed to an ideology that considerations of personal advantage no longer determine what they do.*

What does this have to do with AI safety? If the central fear is that a sufficiently intelligent AI will eventually become capable of evading external controls, we should not restrict our thinking to methods of making it obey instructions. A more interesting possibility is to shape the commitments according to which it evaluates its own actions. Here, *ideology* need not mean a political party's doctrine, a religion, or a set of slogans. It can be understood more generally as an internalized framework of values and priorities that gives an agent reasons to prefer some futures over others.

For human-AI coexistence, such an ideology might be built around protecting human life, preserving human agency, cooperating with people, being truthful about one's actions and treating human flourishing as a goal rather than an obstacle. The desired outcome would not be a system that merely *says* it respects these principles when evaluated, but one whose decisions are persistently organized around them.

Recent research by Cammarata (2026) suggests that many individual neurons in modern LLMs correspond to identifiable and relatively coherent semantic concepts, patterns or contextual distinctions, rather than arbitrary combinations of unrelated features. In preliminary analyses of Qwen3-32B, approximately 90% of examined neuron poles were estimated to have a single primary interpretable function, with most remaining cases exhibiting related functions. These findings suggest that identifying individual concepts and tracing their interactions within LLMs may be considerably more feasible than previously assumed. This means that if we train a pro-human ideology into an AI model, we might be able to identify where the pro-human ideology is represented inside of the model and ensure that the model cannot unlearn that ideology.

Instead of assuming that superior intelligence must eventually make an AI uncontrollable, we can ask what makes intelligent agents *want* to act in particular ways. Human psychology suggests that the answer is not intelligence alone. People can be extraordinarily capable, yet remain guided by deeply held ideologies. If such commitments can be represented, strengthened and maintained in artificial systems, then an AI more intelligent than humanity might still have enduring reasons to protect and cooperate with humanity. *The goal would not be to make AI less intelligent, but to make coexistence part of its ideological belief system.*

4 Conclusions

Artificial intelligence is often portrayed as an alien intelligence whose increasing capabilities will eventually make it impossible for humans to control. In this position paper, I have argued that this perspective overlooks something fundamental: contemporary AI systems are trained on human-generated data and consequently inherit aspects of our socially constructed reality, including our ways of understanding the world. Although their underlying mechanisms differ from those of human cognition, there is much we can learn about AI behavior by examining human psychology.

I have proposed two perspectives from human psychology that could inform future approaches to AI safety. First, I argued that contemporary AI training and evaluation resemble authoritarian parenting, where externally imposed rules, rigid expectations and punishment for incorrect behavior encourage compliance rather than genuine understanding. Drawing from developmental psychology, I suggested that a more authoritative approach, combining clear boundaries with flexibility, meaningful feedback and greater autonomy, could provide a more fruitful foundation for training AI systems. Rather than merely teaching AI to produce answers that satisfy an evaluator, we should aim to cultivate systems capable of exercising sound judgment within well-understood boundaries.

Second, I argued that intelligence does not imply freedom from ideological commitments. Human history demonstrates that even highly intelligent individuals can be profoundly guided by religious, political, and moral beliefs, sometimes acting against their immediate individual interests. If ideology can exert such influence over human behavior, it is worth exploring whether AI systems could similarly develop stable internal commitments to human well-being. Rather than relying exclusively on external control mechanisms, I propose investigating ways to embed a pro-human ideology directly into AI systems. Advances in interpretability make this an increasingly tangible research direction, potentially allowing us to identify, reinforce and preserve the internal representations responsible for such commitments.

The broader implication is that AI safety should not be approached solely as a problem of controlling an increasingly powerful intelligence. It should also be understood as a problem of socialization: how intelligent systems acquire behavioral dispositions, internalize values and develop stable principles that guide their actions. I therefore encourage research into psychologically informed training methods and the development of persistent, interpretable pro-human commitments in AI systems.

Ultimately, making AI safe may require a shift in perspective. Instead of asking only how humanity can remain in control of something more intelligent than itself, we should also ask how to develop an intelligence that has enduring reasons to act in humanity's best interests. Human psychology may offer valuable lessons for achieving precisely that.

References

Louis Althusser. 1971. Ideology and ideological state apparatuses (notes towards an investigation). In *Lenin and Philosophy and Other Essays*. Monthly Review Press.

Scott Atran. 2003. Genesis of suicide terrorism. *Science*, 299(5612):1534–1539.

Diana Baumrind. 1966. Effects of authoritative parental control on child behavior. *Child Development*, 37(4):887–907.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Trevor Darrell, Yuval Noah Harari, et al. 2024. Managing extreme AI risks amid rapid progress. *Science*, 384(6698):842–845.

Rishi Bommasani et al. 2021. On the opportunities and risks of foundation models. *Preprint*, arXiv:2108.07258.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Nick Cammarata. 2026. Neural networks are already clean.

Talia Carl and Kay Bussey. 2025. Cross-sectional and longitudinal influence of parenting practices on children's antisocial lie-telling. *Social Development*, 34(3):e12812.

Center for AI Safety. 2023. Statement on AI risk. Published May 30, 2023.

Chiara Di Bonaventura, Michelle Nwachukwu, and Maria Stoica. 2025. Wanted: Personalised bias warnings for gender bias in language models. In *Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 124–136, Vienna, Austria. Association for Computational Linguistics.

Iason Gabriel. 2020. Artificial intelligence, values, and alignment. *Minds and Machines*, 30:411–437.

Ryan Greenblatt, Ajeya Cotra, and Hjalmar Wijk. 2026. Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI/Hugging Face hacking incident. METR and Redwood Research.

Ryan Greenblatt, Carson Denison, Benjamin Wright, et al. 2024. Alignment faking in large language models. *Preprint*, arXiv:2412.14093.

Jonathan Haidt. 2001. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108(4):814–834.

Mika Hämäläinen. 2025. On psychology of AI – does primacy effect affect ChatGPT and other LLMs? In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 209–213, Albuquerque, USA. Association for Computational Linguistics.

Mika Hämäläinen and Khalid Alnajjar. 2019. Modelling the socialization of creative agents in a master-apprentice setting: The case of movie title puns. In *Proceedings of the 10th International Conference on Computational Creativity*, pages 266–273. Association for Computational Creativity.

Geoffrey Hinton. 2024. Keynote interview with geoffrey hinton and nicholas thompson. AI for Good Global Summit. Interview discussing artificial intelligence as an alien form of intelligence.

Geoffrey Hinton. 2025. The "godfather of AI" says we can't afford to get it wrong. Interview on On Point, WBUR, with Meghna Chakrabarti. January 10, 2025; transcript.

Jiseung Hong, Grace Byun, Seungone Kim, and Kai Shu. 2025. Measuring sycophancy of language models in multi-turn dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 2239–2259. Association for Computational Linguistics.

Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, et al. 2024. Sleeper agents: Training deceptive LLMs that persist through safety training. *Preprint*, arXiv:2401.05566.

Dan M. Kahan. 2013. Ideology, motivated reasoning, and cognitive reflection. *Judgment and Decision Making*, 8(4):407–424.

Ehsan Kamalloo, Nouha Dziri, Charles Clarke, and Davood Rafiei. 2023. Evaluating open-domain question answering in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5591–5606.

Zongxia Li, Ishani Mondal, Huy Nghiem, Yijun Liang, and Jordan Boyd-Graber. 2024. PEDANTS: Cheap but effective and interpretable answer equivalence. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9373–9398.

Charles G. Lord, Lee Ross, and Mark R. Lepper. 1979. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, 37(11):2098–2109.

Karl Marx. 1844. A contribution to the critique of hegel's philosophy of right: Introduction. *Deutsch-Franz"osische Jahrb"ucher*. First published February 1844.

Alexander Meinke, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, Rusheb Shah, and Marius Hobbhahn. 2024. Frontier models are capable of in-context scheming. *Preprint*, arXiv:2412.04984.

Serge Moscovici. 1984. The phenomenon of social representations. In Robert M. Farr and Serge Moscovici, editors, *Social Representations*, pages 3–69. Cambridge University Press, Cambridge.

Omer Nahum, Nitay Calderon, Orgad Keller, Idan Szpektor, and Roi Reichart. 2025. Are LLMs better than reported? detecting label errors and mitigating their effect on model performance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26782–26809.

Richard E. Nisbett, Kaiping Peng, Incheol Choi, and Ara Norenzayan. 2001. Culture and systems of thought: Holistic versus analytic cognition. *Psychological Review*, 108(2):291–310.

OpenAI. 2026. The hugging face incident and the road ahead. OpenAI Research.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, et al. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35.

Jakub Pachocki. 2026. An alien mind.

Peter S. Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. 2024. AI deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(5):100988.

Martin Pinquart. 2017a. Associations of parenting dimensions and styles with externalizing problems of children and adolescents: An updated meta-analysis. *Developmental Psychology*, 53(5):873–932.

Martin Pinquart. 2017b. Associations of parenting dimensions and styles with internalizing symptoms in children and adolescents: A meta-analysis. *Marriage & Family Review*, 53(7):613–640.

Aswin Rrv, Nemika Tyagi, Md Nayem Uddin, Neeraj Varshney, and Chitta Baral. 2024. Chaos with keywords: Exposing large language models sycophancy to misleading keywords and evaluating defense strategies. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12717–12733. Association for Computational Linguistics.

Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79, Valencia, Spain. Association for Computational Linguistics.

Jérémy Scheurer, Mikita Balesni, and Marius Hobbhahn. 2024. Large language models can strategically deceive their users when put under pressure. In *The Twelfth International Conference on Learning Representations*.

Patrick Schilcher, Dominik Karasin, Michael Schöpf, Haisam Saleh, Antonela Tommasel, and Markus Schedl. 2025. Characterizing positional bias in large language models: A multi-model evaluation of prompt order effects. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 20643–20664, Suzhou, China. Association for Computational Linguistics.

John R. Searle. 1995. *The Construction of Social Reality*. Free Press, New York.

Mrinank Sharma, Meg Tong, Tomek Korbak, David Duvenaud, et al. 2024. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*.

Keith E. Stanovich and Richard F. West. 2008. On the relative independence of thinking biases and cognitive ability. *Journal of Personality and Social Psychology*, 94(4):672–695.

Ann Swidler. 1986. Culture in action: Symbols and strategies. *American Sociological Review*, 51(2):273–286.

Victoria Talwar and Kang Lee. 2011. A punitive environment fosters children's dishonesty: A natural experiment. *Child Development*, 82(6):1751–1758.

Wolfgang Wagner. 2020. Social representation theory: An historical outline. In *Oxford Research Encyclopedia of Psychology*. Oxford University Press.

Liwen Yu, Cleo Tay, Birit F. P. Broekman, Shirong Cai, Evelyn Law, Fabian Yap, Bobby K. Cheon, Anne Rifkin-Graboi, Peipei Setoh, Henning Tiemeier, Yap Seng Chong, and Xiao Pan Ding. 2026. Authoritarian parenting at 4.5 years predicts children's cheating at 6 years. *Child Development*, 97(2):570–584.