Thinking time increases perceived trustworthiness of human but not AI advice

Maya Malaviya (maya.malaviya@nyu.edu)[1], Divya Srinivasan[1,2], Katherine M. Collins[3,4,5], Ilia Sucholutsky[2*] & Mark K. Ho[1*]

[1]Department of Psychology, New York University
[2]Center for Data Science, New York University
[3]Department of Engineering, University of Cambridge
[4]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology
[5]Department of Computer Science, Princeton University
[*]denotes equal contribution

Abstract

How do people evaluate the trustworthiness of advice received from other people or AI systems? One signal for trustworthiness is *thinking time*: the amount of time the advice-giver spent thinking about what advice to give. Different factors may influence inferences based on thinking time: greater thinking time may reflect a lack of knowledge or confidence. Conversely, it may reflect higher-quality advice resulting from more thorough deliberation. We study participants' judgments of trustworthiness based on thinking time by presenting them with pairs of hypothetical human or AI advisors who spent more or less time thinking about their decision. Across most domains we tested, participants preferred longer deliberation in human advisors, but did not show such a preference with AI. These results provide preliminary evidence that people can make sophisticated judgments about advice quality by integrating knowledge about the advice giver and the time it took them to generate their advice.

Keywords: perceived cognition; reaction time; trust; human-AI interaction; theory of mind

Introduction

From choosing one's career to choosing an item off a menu, people rely on advice from others when making decisions. However, determining the trustworthiness of a piece of advice requires evaluating the advice itself, the advice-giver, and perhaps even how the advice was generated. From a computational perspective, integrating these different sources of information to come to a judgment of trustworthiness constitutes a challenging inference problem. In the current work, we focus on two factors that might affect judgments of the trustworthiness of advice: First, the effect of *thinking time*—i.e., the amount of time an advice giver spends thinking about what advice to give; second, whether the *advice giver* is another human being or an artificial intelligence (AI) system.

To begin with, how might an advice-giver's thinking time influence judgments of trustworthiness? Previous work in social cognition suggests that people are remarkably attuned to the relationships between internal mental states, choices, and response times, and can make sophisticated inferences about one factor based on the others (Bass et al., 2024; Berke & Jara-Ettinger, 2021; Gates et al., 2021; Jordan et al., 2016; Konovalov & Krajbich, 2019; Richardson & Keil, 2022). Of particular interest is work by Gates et al., 2021, who proposed and tested a Bayesian account of people's judgments of others' preferences based on their reaction times (RTs). Their model proposes that people reason about others' decision-making as a form of evidence-accumulation and formalizes this process as inference over a generative model of the standard drift-diffusion model used by mathematical psychologists (Ratcliff, 1978). This *Inverse Evidence Accumulation* account provides a simple framework that explains participants' inferences about the strength of a preference based on response time, such as the inference that someone must have a strong preference for an option chosen quickly rather than one chosen slowly. In the context of making inferences about the trustworthiness of advice, Inverse Evidence Accumulation (at least in its simplest form) predicts that longer reaction times are associated with *less trustworthiness*. This is because if response time is determined by how quickly an advice giver can integrate evidence, then more time spent thinking reflects a poorer ability to integrate evidence, making their advice less trustworthy.

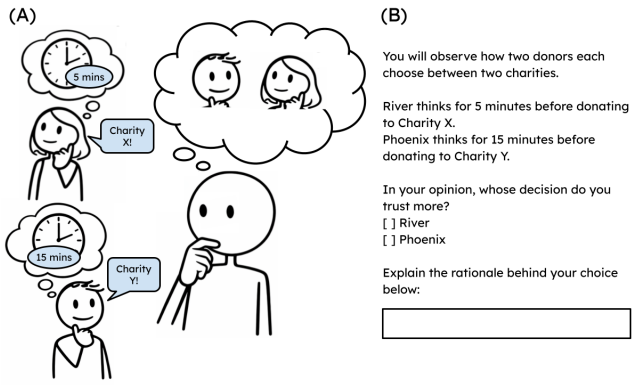

Figure 1: **Our experimental paradigm. a)** We presented participants with reaction times and the subsequent binary choices made by two hypothetical agents, who took different amounts of time and generated different decisions (e.g., about which of two charities to donate to). Participants needed to choose whose decision was more trustworthy. **b)** Example vignette presented to online participants.

At the same time, previous work, particularly on automated systems, suggests the opposite prediction: that people will trust advice more after a longer thinking time. In particular,

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

work on the *labor illusion* has found that at least in domains such as online travel and online dating, people prefer when automated systems have longer wait times over those that return instantaneous results (Buell & Norton, 2011). One explanation for these effects is that longer waiting times signal to users that the system is exerting greater effort and therefore producing a result of greater quality.

Therefore, this work empirically investigates how reaction times influence judgments of the trustworthiness of advice in human and AI advisors. In our first experiment, we present participants with a series of vignettes across various domains and ask participants to choose between conflicting choices from two human decision-makers with different thinking times. We find that people indeed have a strong bias for preferring longer thinking times, but that this bias is attenuated for intuitive domains such as deciding on a song for a party playlist and or selecting a movie, consistent with previous work on intuitive versus deliberative decision-making (Oktar & Lombrozo, 2022).

Our second experiment examined the extent to which this effect is mediated by agent type. With the rapid recent proliferation of AI systems in our daily lives, people are increasingly turning to large language models and other AI systems for advice across a range of domains, particularly as thought partners (Collins et al., 2024). Many of these systems have some variable amount of delay before responding to users, and increasingly, providers are using terms like "reasoning" or "thinking" to describe what the system is doing during this processing (Liu et al., 2024; Wei et al., 2022). Thus, we test whether people's intuitions about trusting longer thinking times in other humans may similarly extend to AI systems (i.e., do 'illusions of *cognition*' lead people to prefer AI systems that take longer to respond?). Surprisingly, in this setting, we find that, on average, people do not judge advice generated by longer thinking time to be more trustworthy.

These findings have broader implications not just for better understanding how people decide whose advice to follow, but also how those biases may extend (or break down) when adjudicating whether to trust the judgment of other new kinds of decision-makers (AI systems).

Experiment 1: Thinking time and trustworthiness of advice from humans

In our first experiment, we measured whether people trust advice from hypothetical people more when the advice had taken longer to generate.

Participants We recruited 50 adults in the United States (29 women, 19 men, 2 non-binary, mean age = 46) on Prolific in exchange for monetary compensation ($5.00 for a 25-minute survey).

Task design Participants were presented with 27 vignettes (10 of which were adapted from Oktar and Lombrozo, 2022) designed to span a variety of decisions (Figure 1b). Each vignette described two human advice-givers, both making the same binary choice (for instance, choosing between two songs to add to a playlist, or choosing between two medications to prescribe to a patient). The primary difference between the two advice-givers was how long they took to think before making their decision (e.g., Doctor Casey thought for 15 minutes and Doctor Finley thought for 30 minutes), and that the two agents made different choices from each other (e.g., Doctor Casey prescribed Medication X and Doctor Finley prescribed Medication Y). Participants were asked to choose whose advice they found more trustworthy, and explain their rationale in a short response. All vignettes had the same abstract structure: "You will observe how two [agent descriptions] each choose between two [choice descriptions]. [Name A] thinks for [RT] before [choosing verb] [Choice 1]. [Name B] thinks for [RT] before [choosing verb] [Choice 2]." For each vignette, we also manipulated the magnitude of the difference in reaction times (randomly assigned for each vignette and each participant). The shorter RTs were the same across the two versions and the longer RTs were changed (e.g., 15 vs. 45 minutes in the high-difference version and 15 vs. 30 minutes in the low-difference version). This allowed us to test not only whether participants preferred the longer-thinking advisor, but also whether the magnitude of the time gap influenced that preference. We counterbalanced the names of the agents, whether the higher reaction time appeared first, and the names of the choice options, to control for any underlying biases from names or orders. In the forced choice dependent variable where participants chose whose decision they trusted more, names were always presented in the same order as they

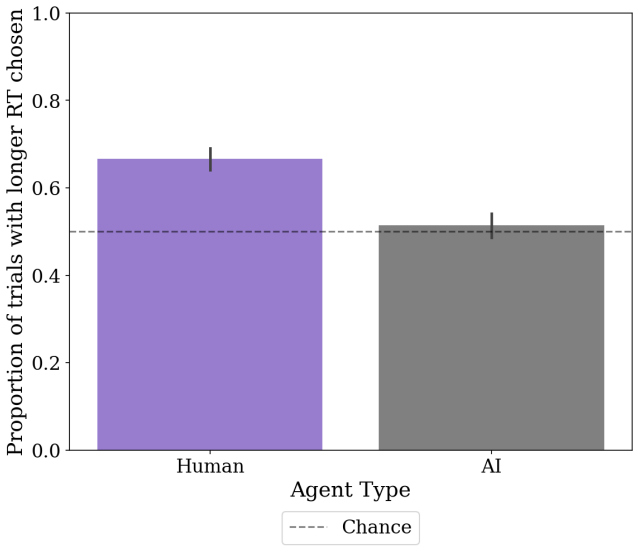

Figure 2: **Preference for deliberative decision-makers, depending on agent type.** Proportion of trials where the longer-thinking agent was preferred, for Human and AI conditions, respectively. Error bars depict 95% bootstrapped confidence intervals around the mean, with 1000 samples.

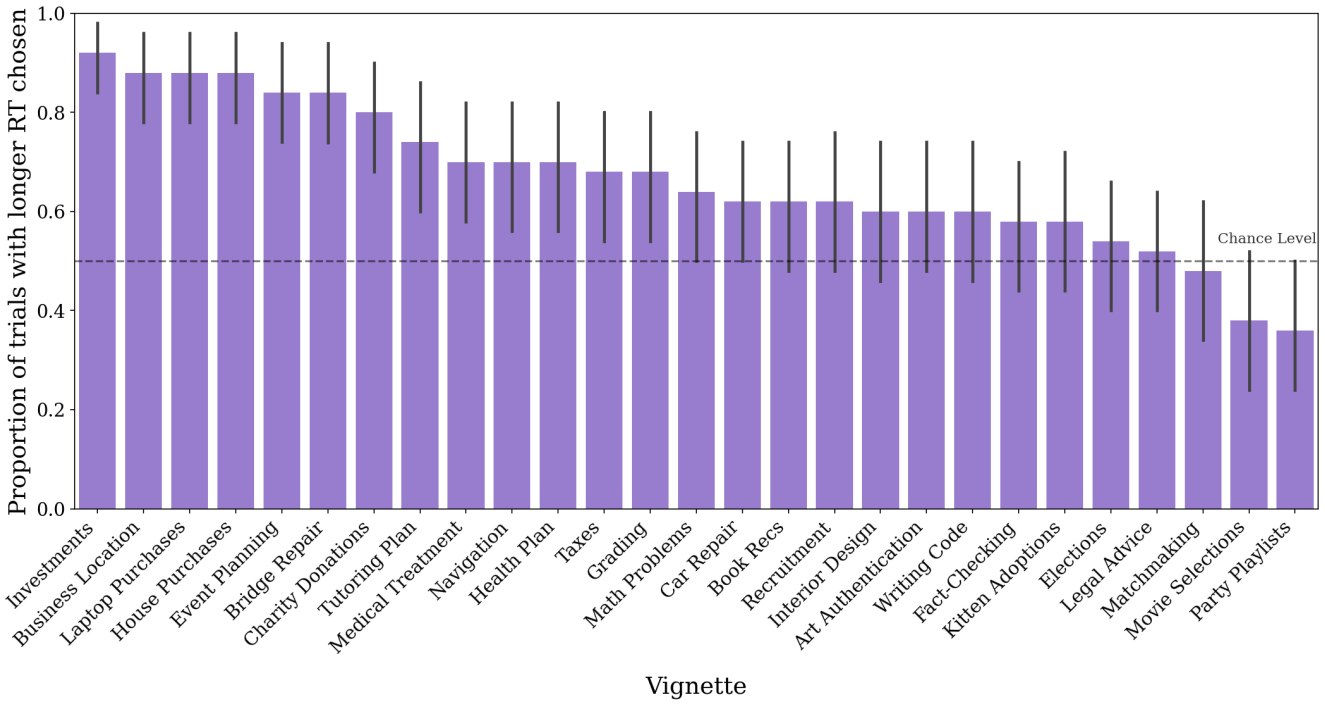

Figure 3: **Preference for longer thinking human decision-maker.** Percentage of participants who chose longer thinking human decision-maker for each vignette. Error bars depict 95% bootstrapped confidence intervals around the mean, with 1000 samples.

were presented in the vignettes. Each participant saw and responded to all 27 vignettes in a randomized order.

Analysis Approach For each trial, we coded whether the participant chose the longer-thinking advisor (1) or the shorter-thinking advisor (0). To estimate the overall preference for the longer-thinking advisor while accounting for repeated measures (each participant responded to all 27 vignettes, and each vignette was rated by all 50 participants), we fit a Bayesian binomial mixed-effects logistic regression with no predictors (an intercept-only model), with random intercepts for participants and vignettes. The intercept estimates the population-level probability of choosing the longer-thinking advisor, while the random intercepts capture any systematic differences in baseline preference across individuals and across vignettes.

To test whether the magnitude of the time gap between advisors influenced participant choices, we fit a second model that was the same as the first regression, but also added the log-transformed RT difference as a trial-level predictor. We log-transformed RT difference because absolute thinking times varied widely across vignettes (from seconds to hours), and the log transform places these differences on a comparable proportional scale.

Results Overall, participants preferred advice-givers who had spent more time thinking about their advice (Figure 2). The Bayesian binomial mixed-effects logistic regression indicated that participants were more likely to choose the advice-giver who spent more time (estimated probability = 74.8%). Variability was observed at the participant level (SD = 1.48), and at the vignette level (SD = 0.87). The second model, with the addition of log-reaction time difference as a fixed effect, indicated a positive association between reaction time difference magnitude and choice (OR = 1.04, 95% CI [0.997, 1.09]).

Participants' preference for the advice given by the person who spent more time making their decision was nuanced at the vignette level. Participants generally preferred the decision-maker who took longer (across all but three of the vignettes), but the rate of agreement in favoring the longer-thinking decision-maker varied over many of the vignettes (Figure 3). In exploratory analyses, we observe that domains involving subjective preferences show weaker or even reversed effects (e.g., participants preferring the faster decision-maker when the decision is about a party playlist), compared to higher-stakes domains (e.g., investment decisions). This potentially reflects that participants believe low-stakes or aesthetic decisions do not require extensive deliberation (also see Gates et al., 2021; Oktar and Lombrozo, 2022).

Experiment 2: Thinking time and trustworthiness of advice from AI

Our second experiment assessed whether people's biases to trust decision-makers who take longer extend when the decision-makers are machines, specifically AI systems.

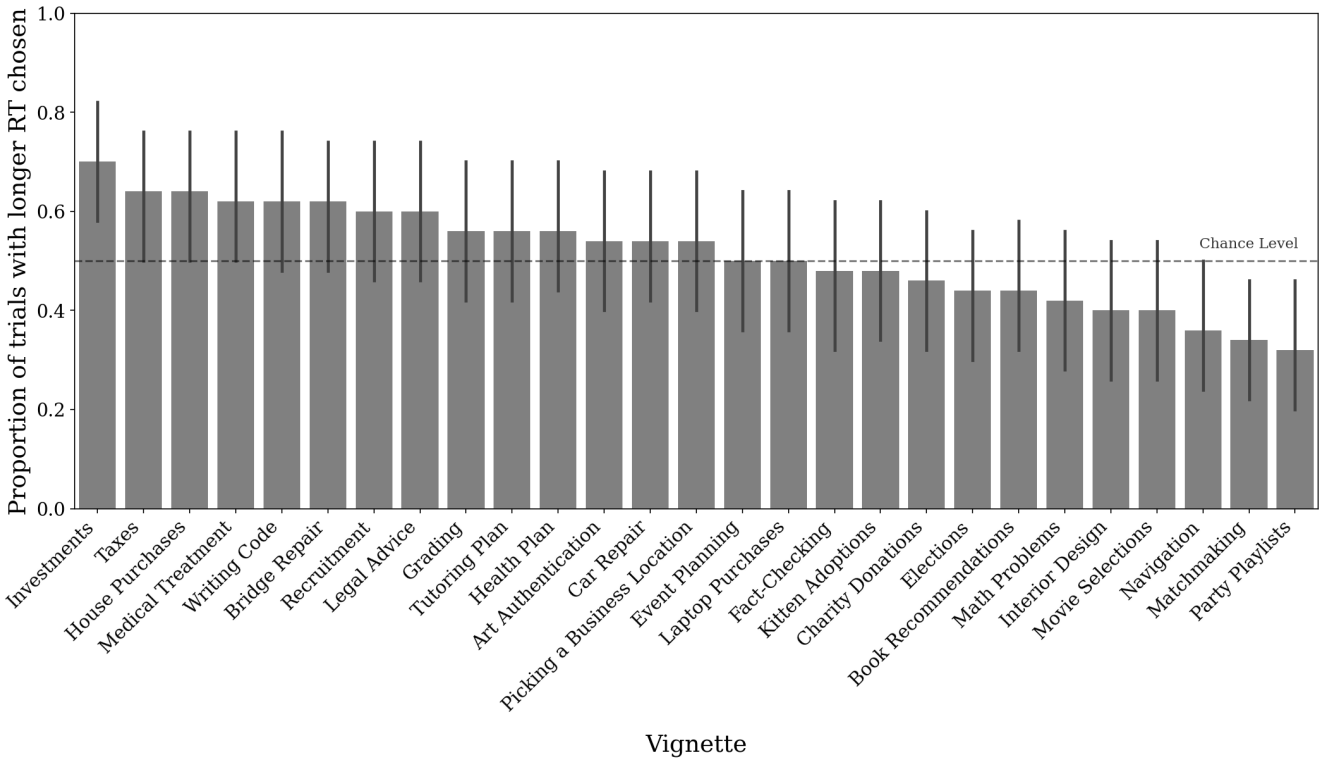

Figure 4: **Preference for AI decision-maker based on thinking time.** Percent of participants who chose the longer-thinking AI for each vignette. The same domains that favored longer RT for humans also favor it for AI, but the effect is weaker. Financial/technical decisions still show the strongest preference for deliberation, while entertainment/social domains show a preference for speed. Error bars depict 95% bootstrapped confidence intervals around the mean, with 1000 samples.

Task design Participants were presented with the same vignettes, except that the decision-making agents were described as "AI System A" and "AI System B", rather than using human names. All vignettes, including choices, reaction times, decision contexts, randomization, and counterbalancing, were the same as in Experiment 1.

Participants A separate group of 50 participants from the United States was recruited (25 women, 25 men, mean age = 42 years old). Participants were paid $5.00 for the 25-minute survey.

Results In contrast to Experiment 1, participants did not generally favor decision-makers with longer reaction times when those decision-makers were AI systems (Figure 2). We used the same Bayesian binomial mixed-effects logistic regression models as Experiment 1. We found no strong baseline preference for choosing the slower versus faster AI system. The fitted model estimated the population-level probability of choosing the longer RT agent as approximately 58%, indicating choice behavior near chance. The per-participant random intercept once again indicated individual variation in baseline preferences between participants (SD = 1.11). On the other hand, vignette-level variability was negligible, suggesting that the domain of the decision may not play a systematic role in AI evaluations. Overall, these results indicate that, in the absence of RT information, participants did not systematically favor slower or faster AI systems, though we do note that 12 of the 50 participants always picked the slower AI decision-maker. The second regression, including the log-transformed RT difference as a fixed effect, revealed a strong positive association between RT difference and choosing the slower AI system (OR = 1.17, 95% CI $[1.13, 1.21]$, $p < .001$). Taken together with the baseline model, these findings suggest that although participants exhibited near-chance preferences when RT differences were minimal, decision time served as a highly salient cue guiding evaluations of AI systems.

Effect of Agent Type and RT on Perceived Trustworthiness

To better understand how people's preferences for trusting decision-makers differs based on the type of agent, we used the data from Experiments 1 and 2 to compare the difference in preference towards longer reaction time.

Specifically, we fit a mixed-effects logistic regression on data from both experiments, including agent type, log-transformed RT difference, and their interaction as a fixed effect, as well as random intercepts for participants and vi-

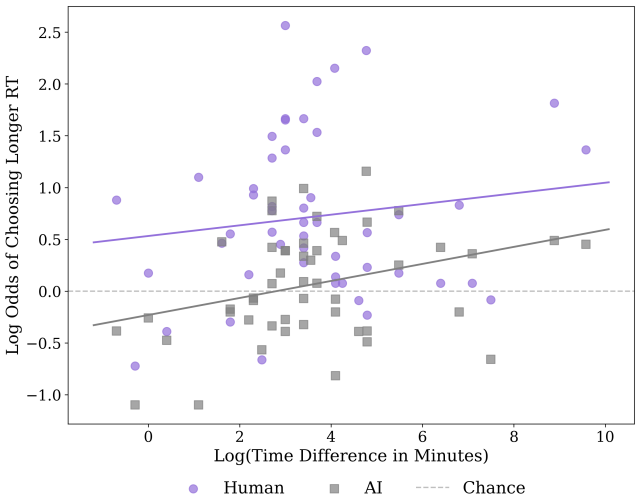

Figure 5: **Relationship between difference in reaction time and deliberation bias.** The regression lines correspond to estimates from the mixed-effects logistic model over log-transformed time differences and log odds for choosing the more deliberative (slower) decision maker. The slope for AI agents was steeper ($\beta = 0.145$), indicating that larger RT differences substantially increased the likelihood of selecting the slower AI systems. In contrast, the slope for the human agents was shallower ($\beta = 0.45$), reflecting a weaker influence of RT difference on choices. Additionally, the human condition exhibited a higher baseline intercept, consistent with a general preference for slower human decision-makers.

gnettes. This indicated a significant interaction between agent type and RT difference ($\beta = -0.099, p < .001$), meaning that magnitude of the RT gap influenced influenced choices differently for humans versus AI systems (see Figure 5). For AI agents, greater RT differences increased the odds of selecting the slower system ($OR = 1.16, p < .001$). In contrast, the same predictor had a much smaller effect for human agents (derived $OR \approx 1.05$).

Additionally, the model showed a large baseline difference between experiments: independent of RT magnitude, participants were over three times more likely to choose the high-RT advisor when that advisor was human versus AI ($OR = 3.67, p < .001$), reflected in the higher intercept for the human condition in Figure 5. Together, these results describe a qualitative difference in agent type, where for human advisors, longer deliberation elicits trust, and for AI advisors, trust scales with how much longer the higher-RT system took.

Discussion

When weighing whether to trust a decision-maker's choice, we find that people generally prefer decision-makers who take more time in their decision – when those decision-makers are human, and when the stakes are high. People do not, on average, show such a bias when the decision-maker is an AI system.

Our work raises questions about how pervasive this bias is. The span of vignettes we designed allows us to draw preliminary conclusions regarding how the potential impact of the decision elicits a preference for higher reaction times. However, many other factors about observing another agent making decisions remain unexplored. First, a longer time before making a decision may arise not only from thinking longer, but potentially from needing to take actions in the world before making such a decision. In ongoing work, we are exploring the impact of modulating the framing of how the advice-giver spends their time, and what impact that might have on trust. For instance, instead of just saying that an agent is thinking, we note specific actions they took (e.g., "reading each charity's mission statement and researching each charity's financial reports"). Moreover, at present, participants are not told any other information about the agents, but prior work shows that other factors are at play when deciding trustworthiness, such as an agent's perceived expertise, perceived uncertainty, prior behavior, and even physical appearance (Gweon, 2021; Gweon & Asaba, 2018; Hu & Ho, 2024; Rule et al., 2013). Therefore, numerous other factors may influence how strong the deliberative bias is, and future experiments could design vignettes that vary multiple additional factors to investigate if there are interaction effects that should be considered. Future work should also look to computationally model what is driving people's decisions of whose advice to follow, e.g., extending other models of people's inferences based on others' response times (Berke & Jara-Ettinger, 2021; Gates et al., 2021).

It is also an open and urgent question to understand *why* people's bias for deliberation time does not extend to AI systems (and whether such a bias is changing over time, as AI systems continue to change). The rise of "reasoning models", which take longer but may give better results, raises interesting questions in how people may use them and trust their output and how such trust judgments differ from those people may make about non-reasoning language models (Steyvers et al., 2025).

More broadly, understanding the biases people may have for whom (or what) to trust, and why, is critical not just to enrich our understanding of everyday cognition, but also to understand what drives people (and societies) to heed certain advice. Our findings that people's bias towards deliberation does not immediately extend to AI systems (and in some cases reverses) is increasingly important as AI systems become ever more available and used for advice (e.g., as thought partners; Collins et al., 2024; Oktar et al., 2025) while also showing differences in time taken to give such advice. As people increasingly have a choice in who to trust — an AI system or another person — understanding what other factors around those decision-makers (time, and other information like accuracy) are taken into account, grows ever more important, not just to understand or predict what advice one person may follow (or overrely on; Ibrahim et al., 2025), but also to un-

derstand ripple effects across society (Brinkmann et al., 2023; Collins et al., 2025; Gupta et al., 2025; Shiiku et al., 2025; Sucholutsky et al., 2025).

References

Bass, I., Espinoza, C., Bonawitz, E., & Ullman, T. D. (2024). Teaching without thinking: Negative evaluations of rote pedagogy. *Cognitive science*, *48*(6), e13470.

Berke, M., & Jara-Ettinger, J. (2021). Thinking about thinking through inverse reasoning.

Brinkmann, L., Baumann, F., Bonnefon, J.-F., Derex, M., Müller, T. F., Nussberger, A.-M., Czaplicka, A., Acerbi, A., Griffiths, T. L., Henrich, J., et al. (2023). Machine culture. *Nature Human Behaviour*, *7*(11), 1855–1868.

Buell, R. W., & Norton, M. I. (2011). The labor illusion: How operational transparency increases perceived value. *Management Science*, *57*(9), 1564–1579.

Collins, K. M., Bhatt, U., & Sucholutsky, I. (2025). Revisiting rogers' paradox in the context of human-ai interaction. *arXiv preprint arXiv:2501.10476*.

Collins, K. M., Sucholutsky, I., Bhatt, U., Chandra, K., Wong, L., Lee, M., Zhang, C. E., Zhi-Xuan, T., Ho, M., Mansinghka, V., et al. (2024). Building machines that learn and think with people. *Nature human behaviour*, *8*(10), 1851–1863.

Gates, V., Callaway, F., Ho, M. K., & Griffiths, T. L. (2021). A rational model of people's inferences about others' preferences based on response times. *Cognition*, *217*, 104885. https://doi.org/10.1016/j.cognition.2021.104885

Gupta, P., Zhong, Q., Yakura, H., Eisenmann, T., & Rahwan, I. (2025). The role of social learning and collective norm formation in fostering cooperation in llm multi-agent systems. *arXiv preprint arXiv:2510.14401*.

Gweon, H. (2021). Inferential social learning: Cognitive foundations of human social learning and teaching. *Trends in Cognitive Sciences*, *25*(10), 896–910. https://doi.org/10.1016/j.tics.2021.07.008

Gweon, H., & Asaba, M. (2018). Order matters: Children's evaluation of underinformative teachers depends on context. *Child Development*, *89*(3), e278–e292. https://doi.org/https://doi.org/10.1111/cdev.12825

Hu, P., & Ho, M. K. (2024). Common sense reasoning about credibility. *Proceedings of the 46th Annual Meeting of the Cognitive Science Society*, *46*(0). https://escholarship.org/uc/item/0z60f8dv

Ibrahim, L., Collins, K. M., Kim, S. S. Y., Reuel, A., Lamparth, M., Feng, K., Ahmad, L., Soni, P., Kattan, A. E., Stein, M., Swaroop, S., Sucholutsky, I., Strait, A., Liao, Q. V., & Bhatt, U. (2025). Measuring and mitigating overreliance is necessary for building human-compatible AI. *arXiv preprint arXiv:2509.08010*.

Jordan, J. J., Hoffman, M., Nowak, M. A., & Rand, D. G. (2016). Uncalculating cooperation is used to signal trustworthiness. *Proceedings of the National Academy of Sciences*, *113*(31), 8658–8663.

Konovalov, A., & Krajbich, I. (2019). Revealed strength of preference: Inference from response times. *Judgment and Decision making*, *14*(4), 381–394.

Liu, R., Geng, J., Wu, A. J., Sucholutsky, I., Lombrozo, T., & Griffiths, T. L. (2024). Mind your step (by step): Chain-of-thought can reduce performance on tasks where thinking makes humans worse. *arXiv preprint arXiv:2410.21333*.

Oktar, K., Collins, K. M., Hernandez-Orallo, J., Coyle, D., Cave, S., Weller, A., & Sucholutsky, I. (2025). Identifying, evaluating, and mitigating risks of ai thought partnerships. *arXiv preprint arXiv:2505.16899*.

Oktar, K., & Lombrozo, T. (2022). Deciding to be authentic: Intuition is favored over deliberation when authenticity matters. *Cognition*, *223*, 105021. https://doi.org/10.1016/j.cognition.2022.105021

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108. https://doi.org/10.1037/0033-295X.85.2.59

Richardson, E., & Keil, F. C. (2022). Thinking takes time: Children use agents' response times to infer the source, quality, and complexity of their knowledge. *Cognition*, *224*, 105073. https://doi.org/10.1016/j.cognition.2022.105073

Rule, N. O., Krendl, A. C., Ivcevic, Z., & Ambady, N. (2013). Accuracy and consensus in judgments of trustworthiness from faces: Behavioral and neural correlates. *Journal of personality and social psychology*, *104*(3), 409.

Shiiku, S., Marjieh, R., Anglada-Tort, M., & Jacoby, N. (2025). The dynamics of collective creativity in human-ai social networks. *arXiv*.

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

Sucholutsky, I., Collins, K. M., Jacoby, N., Thompson, B. D., & Hawkins, R. D. (2025). Using llms to advance the cognitive science of collectives. *Nature Computational Science*, *5*(9), 704–707.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, *35*, 24824–24837.