TANGO

D1.3 Experimental Study of Decision Evaluation Factors

Submission date: 31/03/2026

Due date: 31/03/2026

DOCUMENT SUMMARY INFORMATION

Grant Agreement No	101120763	**Acronym**	TANGO
Full Title	It takes two to tango: a synergistic approach to human-machine decision-making		
Start Date	01/10/2023	**Duration**	48 months
Project type	Research and Innovation Action (RIA)	**Funding programme**	Horizon Europe
Deliverable	D1.3: Experimental study of decision evaluation factors		
Work Package	WP1 –Cognitive foundations of mutual understanding and decision making		
Type	R	**Dissemination Level**	PU
Lead Beneficiary	UPC		
Authors	Wim De Neys Matthieu Raoelison Nicolas Beauvais		
Co-authors			

Funded by the European Union

Reviewers	Nick Chater Simon Myers

DISCLAIMER

COPYRIGHT

DOCUMENT HISTORY

Version	Date	Changes	Contributor(s)
V0.1	03/03/2026		Wim De Neys, Matthieu Raoelison, Nicolas Beauvais
V0.2	17/03/2026	Internal reviewer feedback	Nick Chater, Simon Myers
V1	31/03/2026	Final version for review	Wim De Neys, Matthieu Raoelison, Nicolas Beauvais, Nick Chater, Simon Myers

PROJECT PARTNERS

Partner	Country	Short name
UNIVERSITA DEGLI STUDI DI TRENTO	IT	UNITN
TECHNISCHE UNIVERSITAT DARMSTADT	DE	TUDa
UNIVERSITA DI PISA	IT	UNIPI
CONSIGLIO NAZIONALE DELLE RICERCHE	IT	CNR
SCUOLA NORMALE SUPERIORE	IT	SNS
UNIVERSITE PARIS CITE	FR	UPC
FONDAZIONE BRUNO KESSLER	IT	FBK
CARR COMMUNICATIONS LIMITED	IE	CARR
ISTRAZIVACKO-RAZVOJNI INSTITUT ZA VESTACKU INTELIGENCIJU SRBIJE	RS	IVI
SURGICAL SCIENCE SWEDEN AB	SE	SuS
MARTIN LUTHER UNIVERSITY HALLE-WITTENBERG	DE	MLU
CENTRE FOR EUROPEAN POLICY STUDIES	BE	CEPS
BCAM - BASQUE CENTER FOR APPLIED MATHEMATICS	ES	BCAM
AFLIANT SRL (previously U-HOPPER SRL)	IT	AFL (prev UH)
INTESA SANPAOLO SPA	IT	ISP
28DIGITAL (previously EIT DIGITAL)	BE	EITD
SHARE FONDACIJA	RS	SHARE
A11-INITIATIVE FOR ECONOMIC AND SOCIAL RIGHTS	RS	A11
MINISTARSTVO ZA BRIGU O PORODICI I DEMOGRAFIJU	RS	MFWD
AZIENDA PROVINCIALE PER I SERVIZI SANITARI	IT	APSS
SWANSEA UNIVERSITY	UK	UoS
THE UNIVERSITY OF WARWICK	UK	UoW
CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE	FR	CNRS

Funded by
the European Union

LIST OF ACRONYMS

Acronym	Definition
DoA	Description of Action
EC	European Commission
HaDEA	European Health and Digital Executive Agency
WP	Work Package

Executive Summary

Deliverable D1.3 reports results from Task 1.3 within WP1 that focuses on the development of a folk theory of decision making. The work is trying to pinpoint how information about the intuitive or deliberate nature of an interaction partner's decision is affecting our evaluation of the quality and trust in this decision. The deliverable presents the results from our experimental programme. Across 13 preregistered studies, participants were presented with scenarios describing the decision and decision nature (i.e., more intuitive or deliberate) of an interaction partner and evaluated the competence of the partner. In addition to manipulating the described thinking style of the partner (i.e., fast-intuitive or slow-deliberate), we also manipulated the information about the past performance of the interaction partner (i.e., high or low accuracy or no accuracy information). Results of Study 1 show that participants consistently rated deliberative reasoning as higher quality than intuitive reasoning, even when accuracy information was held constant, and perceived deliberative reasoners as smarter and more trustworthy. Study 2-7 validated these results in a set of control experiments (e.g., confirming the generalization across different languages and countries or with incentives). In Study 8-9 we tested Large Language Models (ChatGPT3.5 and 4) to see whether this preference is reflected in their natural language processing (i.e., present in their training data). Results showed it clearly is: we find the same general preference for deliberation over intuition in each of the models. This tentatively suggests that the general preference for deliberation over intuition is deeply entrenched in human (and machine) language and communication–and that LLMs capture this human folk belief. By using time-pressure and cognitive load to experimentally limit deliberate thinking, Study 10-13 indicated that people's deliberation preference is itself intuitive rather than deliberate in nature and occurs automatically–further underlining its robustness. Taken together, the results help clarify people's folk theory about intuitive and deliberate reasoning and shed light on the underlying cognitive mechanism.

All data and materials are available on the OSF open data repository. The findings have been published in the leading open-access Nature Communications Psychology journal.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

1 Introduction

Deliverable D1.3 reports results from Task 1.3 within WP1 (Strategic Objective S01.2) that focuses on the development of a folk theory of decision making. The key goal is to specify the factors affecting machine and human decision quality and trust evaluations. We approach this issue through the influential lens of dual-process theory that has long conceived thinking as an interplay between fast, intuitive and slower, deliberate thought processes (or System 1 and 2, as they are often referred to, e.g., Kahneman, 2011). While this research has illuminated the mechanics of both systems (De Neys, 2023; Evans & Stanovich, 2013; Pennycook et al., 2015), we lack understanding of how people perceive and value these distinct modes of thought—what might be termed a "folk theory" of fast-and-slow thinking (Bonnefon & Rahwan, 2020). That is, how do we actually perceive intuitive versus deliberate reasoning? Do we want people to rely on intuition or deliberation when they reason? Are we more likely to trust someone's advice if they relied on their intuition, or if they took the time to think things through?

These questions carry profound implications for understanding public trust in our judgment, particularly crucial in an era where we face unprecedented exposure to others' opinions and recommendations (Pennycook, 2023). Beyond human reasoning, insight into our folk beliefs is also relevant for AI development (Bonnefon and Rahwan, 2020). Current approaches aim to enhance Large Language Models by emulating deliberate "System 2" reasoning through chain-of-thought prompts and extended computation time (Hagendorff et al., 2023; Snell et al., 2024; Yax et al., 2024). If our folk beliefs indeed favour deliberation, these developments may not only improve AI accuracy but also public trust in its recommendations—potentially helping to overcome inherent algorithm aversion (Jussupow et al., 2020). Likewise, knowledge about our folk beliefs will help to improve human and AI interaction and trust by aligning recommendations based on the preferred reasoning style.

To start addressing this objective, we ran 13 preregistered studies. Below, we present the overall design and results of the experimental programme.

Deliverable outputs overview and compliance

This deliverable fulfils the Description of Action requirement "data in open data repository and published in journal" as follows:

Table 1. Deliverable compliance overview.

Output	Identifier / link	Notes
Peer-reviewed publication	De Neys, W., & Raoelison, M. (2025). Humans and LLMs rate deliberation as superior to intuition on complex reasoning tasks. Nature Communications Psychology, 3, 141. https://doi.org/10.1038/s44271-025-00320-8	

Open data & materials repository	OSF project: https://osf.io/6pc2e/?view_only=a5beec9676914e82baf0bbf4a10e4b64	Repository includes data, materials, preregistrations and analysis scripts.

2 Objectives and scope

D1.3 aims to (i) identify which factors drive evaluators' judgments of reasoning quality in others, (ii) quantify the relative contribution of perceived reasoning mode (intuition vs deliberation) and past accuracy, and (iii) test whether the observed evaluation preferences require deliberative processing during evaluation. A complementary objective is to test whether AI language models reproduce human evaluation patterns.

3 Work performed and methodology

To start addressing these issues, we ran 13 studies using short vignettes. Each vignette described an individual's reasoning style (either fast intuition or slow deliberation) and past accuracy (high, low, or unspecified). Participants then rated the quality of the individual's reasoning by rating, on 11-point scales, how good and smart they perceived the reasoners to be and whether they would follow their advice. Study 1 introduced this paradigm, while Studies 2-7 tested the robustness of the findings.

In Studies 8-9 we presented the vignettes to Large Lange Models (ChatGPT 3.5 and 4) to test whether they had picked up on the human preferences. Finally, Studies 10-13 tested the nature of humans' folk beliefs. We used time-pressure and load manipulations to experimentally reduce deliberate reasoning during the evaluation process. Since deliberation requires both time and cognitive resources, this allowed us to determine whether people's preference for intuition or deliberation depends itself on intuitive versus deliberate reasoning.

Samples. Human participants were recruited via Prolific. Most studies targeted approximately 240 participants; additional samples included French-speaking participants in France and Indian participants in India to probe generalization across language and culture.

LLM studies. The same vignettes were presented via the OpenAI API to ChatGPT 3.5 (Study 8) and ChatGPT 4 (Study 9), with 240 randomized queries per model to mirror human sampling.

Open science. Study designs were preregistered (AsPredicted); data, materials, preregistrations, and analysis scripts are available via OSF.

4 Results

Study 1: General preference for deliberation over intuition

Correlations among our three rating scales were consistently high and were combined into a single composite preference scale reflecting the overall perceived reasoning quality. Figure 1 shows the results. As the figure indicates, participants preferred reasoners with high accuracy over those with low accuracy, with the unspecified condition in between, indicating that the accuracy manipulation was effective as intended.

More critically, we observed a clear general preference for the deliberate over the intuitive reasoner that was strongest when accuracy information was absent and slightly attenuated in the high- and low-accuracy conditions. Importantly, the preference for deliberation never disappeared or reversed and remained significant across all accuracy conditions. These findings thus indicate a general preference for deliberation over intuition, regardless of accuracy.

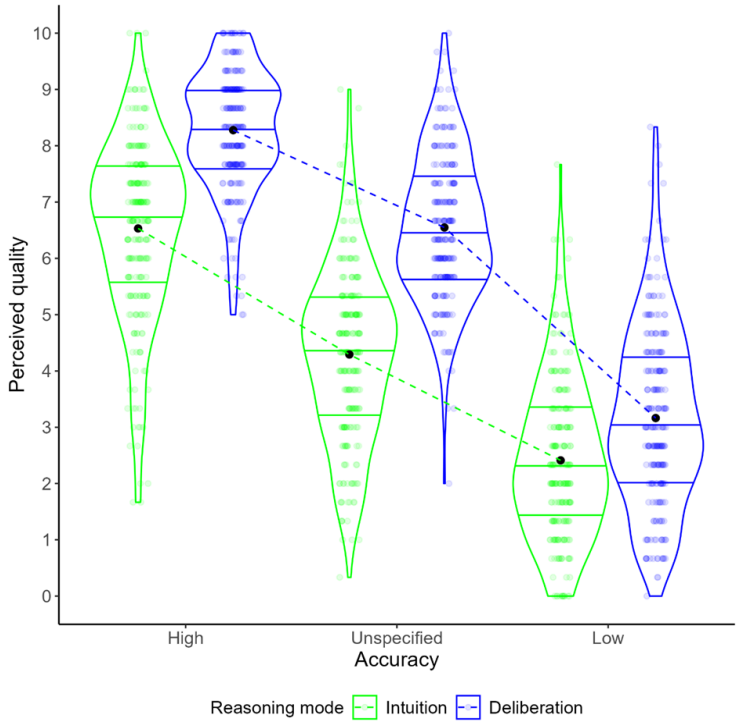

Figure 1. Perceived reasoning quality as a function of an individual's portrayed past accuracy and reasoning mode*. Violin plots showing the overall perceived reasoning quality. Black dots indicate the average. Unsurprisingly, individuals described as having higher past accuracy are preferred over individuals with lower (or unspecified) accuracy. Key result is that individuals who are portrayed as adopting a deliberative reasoning mode (blue) are consistently rated as superior to individuals portrayed as being more intuitive (green), irrespective of past accuracy.*

Study 2-7: Robust preference for deliberation

Study 2-7 served as control studies to validate the robustness of the findings and rule out alternative explanations. Study 2 concerned a simple direct Study 1 replication, while Study 3 and 4 altered vignette content to address specific concerns. One concern was that our verbal accuracy label (e.g., "The accuracy of their answers is very high") might still allow participants to infer that deliberation implicitly yields higher accuracy than intuition, artificially boosting deliberation ratings. To counter this, Study 3 presented a cover story with exact numerical scores on a reasoning test (e.g., "The accuracy of their answers is very high (95%)"), with identical scores for both intuitive and deliberative vignettes.

Another concern was that the intuitive vignette wording ("They do not spend much time or effort to arrive at a conclusion") might skew evaluations negatively by not sufficiently emphasizing the efficiency of fast intuition. In Study 4, we adapted the phrasing to emphasize efficiency (e.g., "They do not need to spend much time or effort to arrive at a conclusion") following pragmatic implicature principles (Grice, 1989). Study 5 addressed potential bias from asking participants for three consecutive ratings by adopting a single evaluation scale for overall reasoning quality.

Participants in Studies 1-5 were all native English speakers living in anglophone, Western countries. Studies 6-7 tested for minimal generalization of the results across languages and cultures. Study 6 tested French-speaking participants in continental Europe, while Study 7 tested Indian nationals living in India.

As shown in Figure 2A, all studies replicated the primary pattern from Study 1. Although accuracy information attenuated the preference, participants consistently favoured deliberation over intuition in all conditions.

Incentivized preference test

At the end of replication Study 2 we also tested whether participants' verbal preference was consequential. Although participants verbally rated deliberation more highly, this preference could potentially stem from social desirability rather than genuine conviction. To test this, we used an incentivized paradigm where participants could double their participation fee by betting on the performance of either a highly accurate intuitor or deliberator (as well as a lowly accurate intuitor and deliberator, for control purposes). A cover story explained that each individual had solved a reasoning test problem, and participants could bet on one of them, receiving a £1 bonus if the individual's answer was correct. In line with the rating findings, participants were far more likely to bet on the highly accurate deliberator (76.8%) than on the intuitor. Interestingly, before the betting question participants also rank ordered the vignettes. Of the participants who ranked the highly accurate deliberator higher than the intuitor, 86.6% also betted on the deliberator. However, among those whose ranking showed a preference for the intuitor, only 46.4% opted for it when money was at stake in the betting question. Thus, when choices had real monetary consequences, the preference for deliberation over intuition became even more pronounced.

Study 8-9: Large Language Models reflect human preferences

In Studies 8-9, we tested whether Large Language Models (LLMs), specifically ChatGPT 3.5 and 4, would mirror the human preference pattern. Using our original vignettes and rating questions, we ran 240 API queries with each model to simulate the same number of subjects as in our individual previous studies. Figure 2B displays the LLM results alongside the aggregate human data from Studies 1–7.

As shown in Figure 2B, both ChatGPT 3.5 and 4 produced response patterns strikingly similar to those of human participants: a consistent preference for deliberation over intuition, strongest when accuracy information was absent and slightly reduced but still significant in the high- and low-accuracy conditions. Notably, the absolute ratings in each condition closely aligned with human responses, with condition means differing on average by 0.62 points for ChatGPT 3.5 and 0.44 points for ChatGPT 4 on the 11-point scale—representing less than 7% deviation from human ratings.

These results suggest that LLMs like ChatGPT have captured human preferences for deliberation over intuition, likely because these patterns are well-represented in the language data used to train the models. This alignment underscores that the preference for deliberative reasoning may be a widely shared belief evident in common language and communication.

Study 10-13: Intuitive nature of deliberation preference

Since Studies 1-9 consistently showed a preference for deliberation over intuition, Studies 10-13 examined the cognitive nature of this preference. One possibility is that the preference for deliberation arises intuitively, without requiring reflection. Alternatively, it might be that individuals actually have an initial intuitive preference for intuition, which shifts towards deliberation when given time to reflect. To test this, we used ever more challenging time-pressure and load manipulations to experimentally reduce deliberate reasoning during the evaluation process in Studies 10-13. If the deliberation preference is itself intuitive, we should observe it even when deliberation is minimized. Conversely, if it requires deliberation, the preference should weaken or reverse when participants are forced to rely on intuition.

Figure 2C presents the results. In each study, light-colored dashed lines indicate conditions where deliberation was experimentally restricted, while dark-colored solid lines indicate unrestricted conditions. As Figure 2C shows, restricting deliberation had minimal impact on the preference for deliberation: participants consistently preferred deliberation over intuition, even under significant constraints, closely matching the patterns seen in Studies 1-7 and the unrestricted conditions in Studies 10-13. This suggests that the preference for deliberation does not depend on deliberate reasoning but is primarily intuitive.

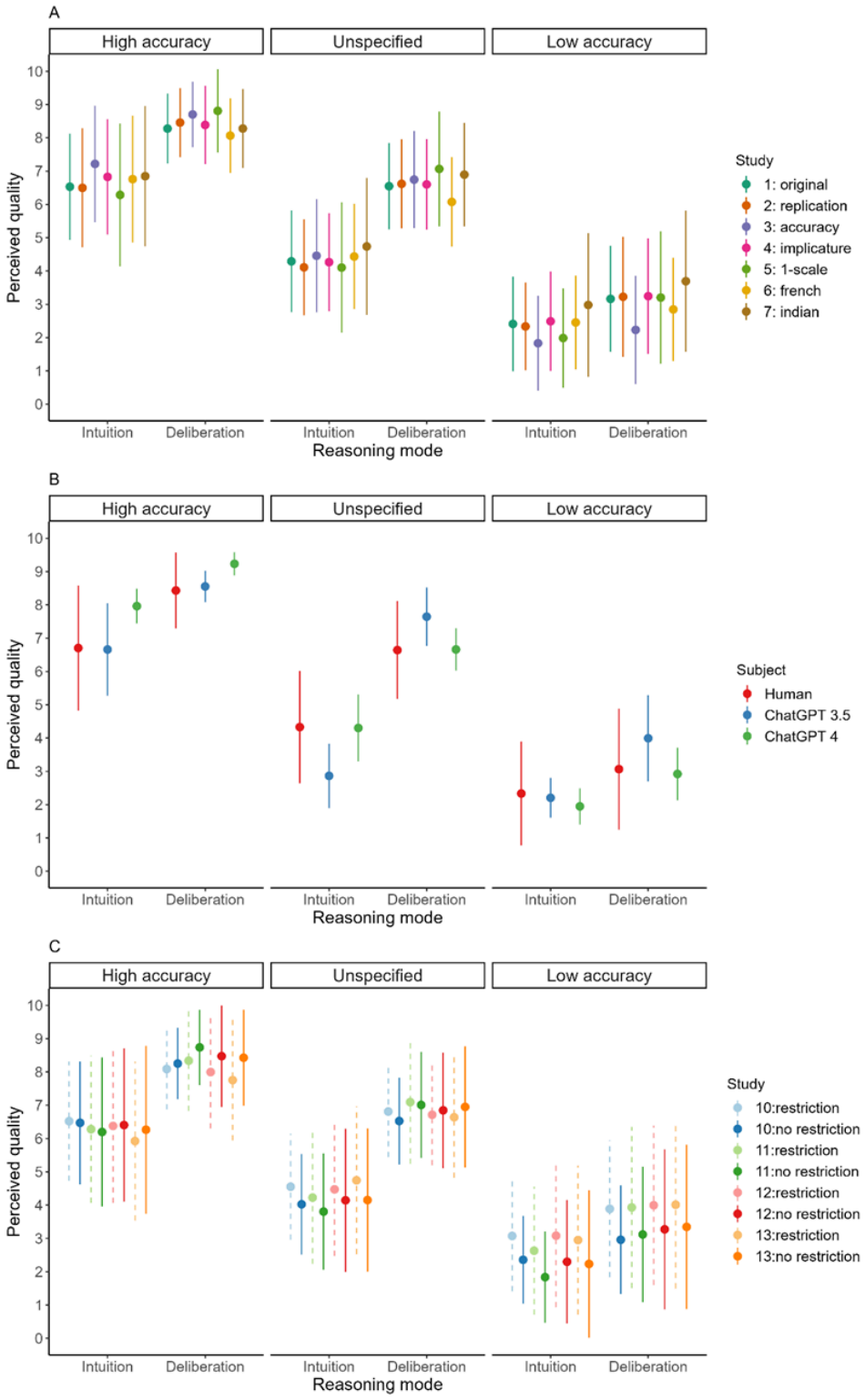

Figure 2. Perceived reasoning quality as a function of an individual's portrayed past accuracy and reasoning mode across all studies*. Panel A shows the results for Studies 1-7 indicating that the general preference for deliberation over intuition is robust and replicated in all control studies. Panel B shows that Large Language Models (Study 8: ChatGPT 3.5; Study 9: ChatGPT 4) reflect the human preferences and also consistently favor deliberation over intuition. Panel C shows the results from Studies 10-13 that*

experimentally restricted deliberation during the evaluation process. Light-colored dashed lines indicate conditions where deliberation was experimentally restricted, while dark-colored solid lines indicate unrestricted conditions. Results indicate that the general preference for deliberation over intuition was still observed under constraints– suggesting it is primarily intuitive in nature.

5 Discussion and project implications

Across 13 studies, we find a robust preference for deliberative over intuitive reasoning, even when accuracy is controlled and participants are under time pressure or cognitive load. Although intuition is often celebrated in popular culture, people seem to readily regard deliberation as a marker of good, thoughtful reasoning. Interestingly, large language models (LLMs) like ChatGPT 3.5 and 4 showed a similar preference, suggesting that the association between deliberation and quality is deeply embedded in human language and communication.

The findings characterise a clear "folk theory" of reasoning: in complex reasoning contexts, deliberation is treated as a credibility cue, linked to higher perceived competence and trustworthiness. This is important for understanding social trust in human judgments and for anticipating how users may evaluate explanations offered by AI systems that explicitly frame their output as deliberative (e.g., "chain-of-thought", extended reasoning, explainability).

In sum, our results help clarify people's folk theory about intuitive and deliberate reasoning and shed light on the underlying cognitive mechanism driving the deliberation preference.

Within the TANGO project, these results can inform subsequent work packages concerned with communication, decision support, and adoption (WP2 and 3): they suggest that recommendations or strategies (e.g., providing explanations, establishing mutual understanding through dialogue, etc.) that are perceived as resulting from extensive thinking may increase perceived reliability and mitigate algorithm aversion. For example, the Explainable and Coherent User Interface (XUI) developed in WP3 (D3.3) turns AI output into visible deliberation. The case study on social welfare (WP5) is set to adopt this specific XUI, which, based on the current results, will increase trust in its recommendations.

But clearly, when considering the broader implications of our findings, it is essential to acknowledge certain limitations of our studies. For example, we focused specifically on folk beliefs about the role of intuition and deliberation in addressing cognitively challenging problems. Our vignettes were designed to depict situations where individuals encounter complex reasoning tasks. More generally, it would be beneficial to investigate the generalizability of these preferences across diverse judgment domains. Related, our analyses focused on the modal reasoners and did not yet try to identify possible individual level variance. These issues will be further tackled in the remainder of the planned WP1-Task 1.3 work (Deliverable D1.4). Finally, while our vignette-based approach allowed us to isolate reasoning mode as a variable, it also comes with limitations. Vignettes, though controlled, do not capture the full ecological complexity of real-world judgments, where factors like social context and personal experience could influence reasoning preferences. In this respect, the realistic TANGO case studies (WP5) will also help to validate the framework further.

6 Data management and ethics considerations

The experimental procedures were approved by the Comité d'Ethique de la Recherche Université Paris Cité (protocol 00012023-151). Participants provided informed consent prior to participation.

The OSF repository provides the datasets and analysis scripts needed to reproduce all reported analyses. Materials include the vignette texts and task instructions. In line with the online testing platform (Prolific Academic) policy we do not record personal identification data.

Recommended citation for reuse: De Neys, W., & Raoelison, M. (2025). Humans and LLMs rate deliberation as superior to intuition on complex reasoning tasks. Communications Psychology, 3, 141. https://doi.org/10.1038/s44271-025-00320-8

7 Next steps

Planned follow-up within the project includes: (i) testing for possible individual-level variance leading to potential divergent preference patterns (ii) testing whether the deliberation preference generalizes to additional judgment domains and real-world decisions.

8 References

Bago, B., & De Neys, W. (2017). Fast logic?: Examining the time course assumption of dual process theory. Cognition, 158, 90-109.

Bago, B., & De Neys, W. (2019). The intuitive greater good: Testing the corrective dual process model of moral cognition. Journal of Experimental Psychology: General, 48, 1782-1801.

Bonnefon, J. F., & Rahwan, I. (2020). Machine thinking, fast and slow. Trends in Cognitive Sciences, 24(12), 1019-1027.

De Neys, W. (2006). Dual processing in reasoning: Two systems but one reasoner. Psychological Science, 17, 428-433.

De Neys, W. (2023). Advancing theorizing about fast-and-slow thinking. Behavioral and Brain Sciences, 46, E111.

Evans, J. St. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition advancing the debate. Perspectives on Psychological Science, 8, 223–241.

Gigerenzer, G. (2007). Gut feelings: the intelligence of the unconscious. New York: Viking.

Gladwell, M. (2006). Blink: The power of thinking without thinking. Boston, MA: Little, Brown, & Company.

Grice, H. P. (1989). Studies in the way of words. Cambridge, MA: Harvard University Press.

Hagendorff, T., Fabi, S., & Kosinski, M. (2023). Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. Nature Computational Science, 3(10), 833-838.

Isler, O., Yilmaz, O. (2023). How to activate intuitive and reflective thinking in behavior research? A comprehensive examination of experimental techniques. Behavior Research Methods, 55, 3679–3698.

Isler, O., Yilmaz, O., & Dogruyol, B. (2020). Activating reflective thinking with decision justification and debiasing training. Judgment and Decision making, 15, 926-938.

Jussupow, E., Benbasat, I., & Heinzl, A. (2020). Why are we averse towards algorithms ? a comprehensive literature review on algorithm aversion. Twenty-Eighth European Conference on Information Systems (ECIS2020), 1–16.

Kahneman, D. (2011). Thinking, Fast and Slow. New York, NY: Farrar, Straus and Giroux.

Kupor, D. M., Tormala, Z. L., Norton, M. I., & Rucker, D. D. (2014). Thought calibration: How thinking just the right amount increases one's influence and appeal. Social Psychological and Personality Science, 5(3), 263-270.

LeGault, M. R. (2007). Think: Why Crucial Decisions Can't Be Made in the Blink of an Eye. New York, NY: Simon & Schuster.

Nye, M., Tessler, M., Tenenbaum, J., & Lake, B. M. (2021). Improving coherence and consistency in neural sequence models with dual-system, neuro-symbolic reasoning. Advances in Neural Information Processing Systems, 34, 25192-25204.

Oktar, K., & Lombrozo, T. (2022). Deciding to be authentic: Intuition is favored over deliberation when authenticity matters. Cognition, 223, 105021.

Pennycook, G. (2023). A framework for understanding reasoning errors: From fake news to climate change and beyond. In Advances in experimental social psychology (Vol. 67, pp. 131-208). Academic Press.

Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2015). What makes us think? A three-stage dual-process model of analytic engagement. Cognitive Psychology, 80, 34–72.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J. F., Breazeal, C., ... & Wellman, M. (2019). Machine behaviour. Nature, 568(7753), 477-486.

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.

Snell, C., Lee, J., Xu, K., & Kumar, A. (2024). Scaling llm test-time compute optimally can be more effective than scaling model parameters. arXiv preprint arXiv:2408.03314.

Swaminathan, V. (2023). Intuition or logic?. Retrieved from https://www.linkedin.com/pulse/intuition-logic-vasudevan-swaminathan/

Thaler, R. H., & Sunstein, C. R. (2021). Nudge: The final edition. Yale University Press.

Umoh, R. (2017). Steve Jobs and Albert Einstein both attributed their extraordinary success to this personality trait. Retrieved from https://www.cnbc.com/2017/06/29/steve-jobs-and-albert-einstein-both-attributed-their-extraordinary-success-to-this-personality-trait.html

Yax, N., Anlló, H., & Palminteri, S. (2024). Studying and improving reasoning in humans and machines. Communications Psychology, 2(1), 51.