Computational Modeling of Heuristic and Base-rate Integration in Reasoning

Jérémie Beucler (jeremie.beucler@gmail.com)
Université Paris Cité, 46 Rue Saint-Jacques, 75005 Paris, France
Zoe Purcell (purcell.z.a@gmail.com)
Université Paris Cité, 46 Rue Saint-Jacques, 75005 Paris, France
Lucie Charles (l.charles@qmul.ac.uk)
Queen Mary University of London, Mile End Road, London E1 4NS, United Kingdom
Kobe Desender (kobe.desender@kuleuven.be)
KU Leuven, Tiensestraat 102, 3000 Leuven, Belgium
Wim De Neys (wim.de-neys@parisdescartes.fr)
CNRS, 46 Rue Saint-Jacques, 75005 Paris, France

Abstract

Base-rate neglect, a key example of biased reasoning, , is often attributed to the interplay between intuitive and deliberative processes in dual-process theories. Yet these explanations remain largely verbal and theoretically underspecified. Participants (N = 151) performed a novel, continuous base-rate neglect task where base-rate information and stereotype-driven heuristic strength (quantified using language models) were parametrically manipulated. Clustering analyses revealed three distinct reasoning profiles: stereotype-driven, base-rate-driven, and balanced. A biased drift diffusion model (DDM), in which weighted stereotype and base-rate information jointly determine the drift rate, captured individual differences and reproduced key empirical patterns in accuracy, confidence, and response time. Results show that confidence and reaction time reflect the same underlying evidence signal as choice, revealing how information is integrated during reasoning. Importantly, the model predicts that biased individuals do not benefit from increased deliberation, as they fail to integrate the logical information in the first place. This work advances the computational modeling of reasoning and offers a theoretical framework for understanding how individuals integrate conflicting information.

Keywords: base rate neglect; reasoning; drift diffusion model; confidence; dual process

Consider the following problem adapted from the classic “lawyer-engineer” problem (Kahneman & Tversky, 1973): A randomly chosen individual from a group of 995 lawyers and 5 engineers is described as "nerdy." Participants are asked: "Which is more likely, that the individual is a lawyer or an engineer?" Despite the explicit base-rate overwhelmingly favoring "lawyer," individuals commonly select "engineer," relying on the stereotype-driven belief rather than on the logical base-rate information. This robust cognitive bias, which significantly impacts real-world decisions, including in medicine (e.g., Bergus et al., 1995) and justice (e.g., Thompson & Schumann, 2017), is known as base-rate neglect.

Dual-process theories explain base-rate neglect, along with similar cognitive biases, through the interplay of two cognitive systems: a rapid, intuitive System 1, and a slower, deliberative System 2 (Evans & Stanovich, 2013; Kahneman, 2011). Recent advances in dual-process theorizing suggest a more nuanced interaction wherein both intuitive and deliberate responses can be activated automatically, with deliberation engagement driven by intuitive assessments of confidence (i.e., “dual-process 2.0”; De Neys, 2023). However, these models remain largely verbal and underspecified, lacking the computational detail needed to formalize how heuristic and probabilistic information are integrated during reasoning.

A key limitation in testing such models has been the lack of systematic methods for quantifying and manipulating heuristic strength. Most studies rely on a binary contrast between conflict items (where stereotype and base-rate information suggest different responses) and no-conflict items (where they align). To overcome this, the present project leverages Large Language Models (LLMs; Le Mens, 2023) to systematically quantify stereotype-driven belief strength.

Participants (N = 151) completed a rapid-response base-rate neglect task (240 trials; adapted from Pennycook et al., 2014), in which LLM-derived stereotype strength varied continuously (e.g., bodybuilders vs. accountants described as “muscular,” “strong,” or “active”). Base-rate ratios were also systematically varied (e.g., 50 bodybuilders/950 accountants, 150 bodybuilders/850 accountants), independently of the stereotype strength. After each response, participants provided a confidence rating. Items were dynamically sampled from an extensive, validated database to ensure balanced coverage across the full range of stereotype strength (Figure 1).

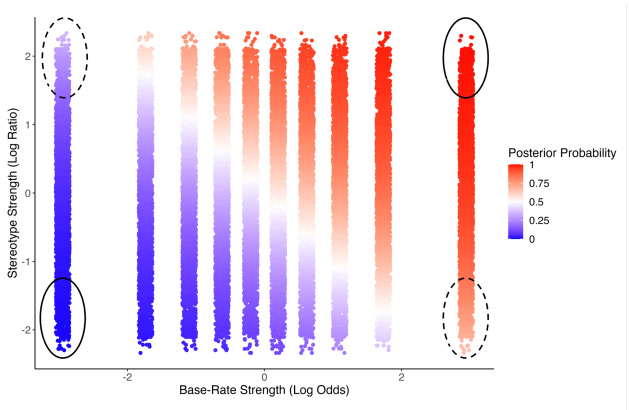

Figure 1: Task space used in the continuous base-rate neglect paradigm. Each dot represents a unique item, plotted as a function of base-rate strength and stereotype strength (log odds/ratios). Color indicates the Bayesian posterior probability assigned to one of the two response categories. Unlike traditional designs that rely only on conflict (dashed ellipses) and no-conflict (solid ellipses) items, the present paradigm samples continuously across the stimulus space.

Behavioral results revealed three distinct clusters based on how participants weighted stereotype versus base-rate information: the majority prioritized stereotypes (54%), a smaller group relied mainly on base-rates (17%), and a third group showed a more balanced use of both (28%). Confidence ratings and reaction times tracked these individual patterns, suggesting metacognitive sensitivity was tightly coupled with participants' initial integration strategy.

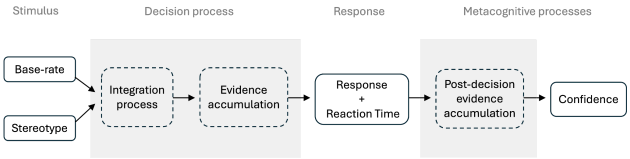

Figure 2: Modeling framework. Stereotype and base-rate information are integrated into a weighted evidence signal, which drives a drift diffusion process leading to a choice. Confidence is computed through post-decision evidence accumulation.

We model participants' choices using a biased Drift Diffusion Model (DDM), where the drift rate is a weighted sum of stereotype and base-rate information, and confidence ratings are derived from post-decisional evidence accumulation (see Figure 2).

Preliminary results suggest that the model effectively captures individual differences in heuristic–base-rate weighting and reproduces key qualitative signatures observed in the empirical data, including patterns of choice, confidence ratings, and reaction times. More specifically, the model mirrors how participants' confidence and response times reflect the relative weighting of heuristic and base-rate information during the decision process.

Our findings illustrate that a single-process biased DDM can successfully account for behavior in a base-rate neglect task by modeling how participants weight stereotype and base-rate information during the reasoning process. Parametrically manipulating both heuristic strength and base-rate information allowed us to capture individual differences in information use and showed that reasoning performance, confidence, and response times all emerge from the same underlying evidence accumulation process.

Our results highlight important theoretical implications: individuals who rely heavily on stereotypes show reduced sensitivity to their reasoning errors, as neither their choices nor their confidence judgments integrate the base-rate information. The model also predicts that increasing deliberation—by raising decision thresholds—does not necessarily improve accuracy in strongly biased reasoners. These findings have implications for cognitive interventions and contribute to refining theoretical models of human reasoning.

References

Bergus, G. R., Chapman, G. B., Gjerde, C., & Elstein, A. S. (1995). Clinical reasoning about new symptoms despite preexisting disease: sources of error and order effects. Family medicine, 27(5), 314–320. https://doi.org/10.1177/0272989X9801800409

De Neys, W. (2023). Advancing theorizing about fast-and-slow thinking. Behavioral and Brain Sciences, 46, e111. doi:10.1017/S0140525X2200142X

Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. Perspectives on Psychological Science, 8(3), 223–241. https://doi.org/10.1177/1745691612460685

Kahneman, D. (2011). Thinking, fast and slow. Penguin Books.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. Psychological Review, 80(4), 237–251. https://doi.org/10.1037/h0034747

Le Mens, G., Kovács, B., Hannan, M. T., & Pros, G. (2023). Uncovering the semantics of concepts using GPT-4. Proceedings of the National Academy of Sciences, 120(49), e2309350120. https://doi.org/10.1073/pnas.2309350120

Pennycook, G., Cheyne, J. A., Barr, N., Koehler, D. J., & Fugelsang, J. A. (2014). Cognitive style and religiosity: The role of conflict detection. Memory & Cognition, 42(1), 1–10. https://doi.org/10.3758/s13421-013-0340-7

Thompson, W. C., & Schumann, E. L. (1987). Interpretation of statistical evidence in criminal trials: The prosecutor's fallacy and the defense attorney's fallacy. Law and Human Behavior, 11(3), 167–187. https://doi.org/10.1007/BF01044641