Hempel's Paradox in Psychological Experiments: Empirical Tests and Large Language Model Simulations

Kuninori Nakamura (knaka@seijo.ac.jp)
Seijo University

Abstract

This paper investigates whether typical experimental practices in cognitive psychology instantiate Hempel's paradox. Using a hypothetical priming experiment, it argues that a failure to observe both concept activation and judgment bias—usually taken as disconfirming—can logically count as confirmation via the contrapositive of the hypothesis. Two studies asked participants to judge whether four possible experimental outcomes confirmed, disconfirmed, or were irrelevant to a hypothesis linking concept activation to biased judgment. Participants consistently failed to treat the contrapositive outcome (no activation and no bias) as confirmatory, both in original and replication contexts. Latent class analyses suggested reliance on satisficing rather than formal logical reasoning. Simulations using a large language model showed partial alignment with human judgments, especially for disconfirmatory outcomes, but clear divergences for confirmatory cases. Overall, the results suggest that standard psychological experiments can give rise to Hempel's paradox and that both humans and LLMs systematically deviate from formal confirmation logic.

Keywords: Hempel's paradox; psychological experiment; hypothesis testing; large language model

Introduction

Imagine the following hypothetical situation. A psychologist proposes the hypothesis that *even when people are not consciously aware of it, activating a certain concept in their minds can bias their judgments and behavior*. To test this hypothesis, the researcher designs the following experiment.

First, participants are shown a particular figure. Based on previous studies, this figure is assumed to be related to the target concept and therefore capable of activating it. After viewing the figure, participants complete two tasks. The first is an association task designed to assess whether the concept has in fact been activated by the figure. If participants recall words related to the concept, this is taken as evidence of concept activation. The second task is a judgment task, in which participants make judgments about a specified target. This task is intended to test whether participants' judgments are biased. If the concept has been activated, its influence should be reflected in their judgments.

Suppose that the results show no evidence of concept activation in the association task and no bias in judgments in the judgment task. Based on these findings, the psychologist concludes that the hypothesis is not supported.

As can be seen, this example represents a typical explanation and experimental design in cognitive psychology, such as those used in research on the selective accessibility hypothesis in anchoring (Mussweiler & Strack, 1999; Tversky & Kahneman, 1974) or on semantic priming effects (Bargh, Chen, & Burrows, 1996). These studies assume that biases in judgment or changes in behavior (e.g., walking speed) are caused by the activation of conceptual knowledge related to the anchor magnitude (Mussweiler & Strack, 1999, 2000) or to elderly stereotypes (Bargh et al., 1996). To test this assumption, researchers attempted to manipulate the activation of conceptual knowledge by presenting participants with relevant words or stimuli. In addition, some studies employed lexical decision tasks to assess whether the relevant conceptual knowledge had actually been activated (Mussweiler & Strack, 2000). If responses to concept-related words were faster in the primed condition than in the unprimed condition, the conceptual knowledge was assumed to have been activated. In this sense, the hypothetical example above closely mirrors actual experimental practice in cognitive psychology.

How, then, should the results in the hypothetical example be interpreted? Intuitively, the researcher's conclusion appears natural and reasonable. However, on closer inspection, this conclusion is logically invalid. The hypothesis entails that activation of the concept leads to biased judgments. Thus, if the association task indicates concept activation and the judgment task shows bias, the hypothesis is supported. At the same time, from a logical perspective, the hypothesis *judgments become biased when a concept is activated* is logically equivalent to the contrapositive claim that *if judgments are not biased, then the concept was not activated*. Therefore, the experimental results in the hypothetical example—showing neither activation nor bias—should also be considered evidence for the hypothesis.

This conclusion is counterintuitive, as it implies that researchers could obtain evidence for the hypothesis even from poorly executed experiments that fail to detect any effects in either task. Nevertheless, from a strictly logical standpoint, the reasoning is valid.

This line of argument illustrates the well-known Hempel's paradox (Hempel, 1945) in the context of a typical experimental situation in cognitive psychology. Hempel's paradox arises from principles of confirmation theory in the philosophy of science. It is based on the logical equivalence between the statements *All ravens are black* and *All non-black things are non-ravens*. According to standard confirmation theory, any observation consistent with a hypothesis provides some degree of confirmation. Consequently, not only observing a black raven but also observing a green apple (a non-black non-raven) should confirm the hypothesis that all ravens are black, because it is logically equivalent to the latter formulation. This leads to the

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

counterintuitive conclusion that seemingly irrelevant observations can count as evidence for a hypothesis.

Now consider a different scenario. Suppose the experiment described above is conducted and the researcher finds evidence that the concept was activated and that judgments were biased. The researcher therefore concludes that the hypothesis is supported. A second researcher, however, questions the credibility of these findings and attempts to replicate the experiment using exactly the same procedure. In the replication study, participants complete the same association task and make the same judgments about the same target. This time, however, the results fail to reproduce the original findings: the association task shows no evidence of concept activation, and the judgment task shows no bias. Paradoxically, although the replication fails to reproduce the original effects, the results still appear—by the same logical reasoning—to support the original hypothesis. We refer to this phenomenon as the *paradox of replication*.

These arguments suggest that typical explanations and experimental practices in cognitive psychology may be vulnerable to Hempel's paradox. Of course, the foregoing discussion is a thought experiment, and how people actually interpret such evidence is an empirical question. Nevertheless, existing research on human reasoning suggests that this prediction is plausible. Numerous studies have shown that people prioritize confirming evidence while underweighting the absence of disconfirming evidence. Research using the number-sequence task (Wason, 1960) and the four-card selection task (Wason, 1966) has consistently demonstrated difficulties in considering disconfirming cases during hypothesis testing. Studies on matching bias (Evans & Lynch, 1973) show that when evaluating conditional statements (if *p*, then *q*), people tend to focus on instances corresponding to *p* and *q*, while neglecting instances corresponding to not-*p* or not-*q*. Similarly, research on covariation assessment indicates that people prioritize observations of joint presence and underweight joint absence when judging the relationship between two variables (e.g., Levin, Wasserman, & Kao, 1993; Kao & Wasserman, 1993; Lipe, 1990; Schustack & Sternberg, 1981; McKenzie & Mikkelsen, 2000; Wasserman, Dorner, & Kao, 1990; for a review, see McKenzie & Mikkelsen, 2008).

Among these studies, McKenzie and Mikkelsen's (2000) work is particularly relevant to the psychological aspects of Hempel's paradox. Participants evaluated a hypothesis concerning the relation between two binary variables—genotype (A vs. B) and personality (X vs. Y)—such as *if a person has genotype A, then they have personality type X*. In this task, both observing *A and X* and observing *B and Y* logically support the hypothesis. Nevertheless, participants judged *A and X* as more supportive than *B and Y*. Furthermore, when information about the base rates of the variables was provided (e.g., genotype A and personality type X being more common than B and Y), participants' preferences were further biased toward observing *A and X*. These results suggest that sensitivity to event rarity plays a role in how people resolve situations related to Hempel's paradox.

In summary, prior research indicates that people tend not to treat contrapositive instances as evidence for a hypothesis, and this tendency may persist even in the context of psychological experimentation. However, to the best of our knowledge, no study has directly examined whether this issue can be understood explicitly in terms of Hempel's paradox within an experimental psychological framework. Accordingly, the present study aims to investigate whether situations analogous to Hempel's paradox arise in psychological experiments.

To address this question, we conducted two empirical studies in which participants evaluated whether different possible outcomes of a psychological experiment would confirm a given hypothesis: *if a certain concept in a person's mind is activated, their judgments will be biased*. After reading a scenario describing such an experiment, participants were presented with four logically possible outcomes: (a) the concept was activated and judgments were biased, (b) the concept was not activated and judgments were biased, (c) the concept was activated and judgments were not biased, and (d) the concept was not activated and judgments were not biased. For each outcome, participants judged whether it confirmed, disconfirmed, or was irrelevant to the hypothesis using a three-alternative forced-choice task.

The procedures of the two studies were identical except for whether the scenario described an original experiment or a replication. Study 1 depicted a researcher proposing and testing an original hypothesis, whereas Study 2 depicted a researcher attempting to replicate another researcher's experiment using the same procedure.

Examination by Large Language Model

In addition to examining how human participants evaluate experimental outcomes, the present study also investigates how a large language model (LLM) would predict participants' judgments in the two studies described above. Recent advances in generative artificial intelligence (AI) have made it possible to explore how AI systems respond to reasoning tasks commonly used in psychological experiments, thereby offering new perspectives on similarities and differences between human and artificial cognition (Binz et al., 2023; Lampinen et al., 2024).

For example, Binz et al. (2023) required GPT-3 to solve vignette-based experimental tasks involving decision making (Kahneman & Tversky, 1979; Tversky & Kahneman, 1983), information search (Hertwig et al., 2004), and causal reasoning (Waldmann & Hagmayer, 2005). They found that the model performed comparably to, or in some cases better than, human participants, while also showing that small perturbations to the task descriptions could lead the model to dramatically different responses. Similarly, Lampinen et al. (2024) demonstrated that LLMs exhibit content effects in Wason's selection task that resemble those observed in human participants, although important divergences between human and model behavior were also observed in certain conditions.

Taken together, these findings suggest that LLMs display systematic patterns of judgment that can be meaningfully compared with those of humans. Such comparisons provide a valuable opportunity to gain insight into the cognitive processes underlying both human reasoning and AI-generated responses. Accordingly, the present study also examines how an LLM evaluates the same experimental scenarios used in Studies 1 and 2, focusing on whether the model's judgments reflect the same asymmetries in confirmation and disconfirmation that characterize human reasoning in situations analogous to Hempel's paradox.

Study 1

Study 1 examined whether naïve participants would evaluate the outcomes of a fictitious psychological experiment in a manner consistent with the implications of Hempel's paradox. Specifically, participants were asked to judge whether each of the four logically possible experimental outcomes described above would confirm the hypothesis. This allowed us to assess whether participants treat contrapositive outcomes—namely, cases in which the concept was not activated and judgments were not biased—as confirmatory evidence for the hypothesis.

Method

Two hundred and ninety-three undergraduate students participated in Study 1. All participants read a vignette describing a situation in which a researcher proposed an original hypothesis and conducted an experiment to test it. After reading the vignette, participants completed a three-alternative forced-choice task, as described in the Introduction. Details of the vignette and the experimental materials are available on the Open Science Framework (OSF; https://osf.io/u9tgp/).

Results and discussion

Figure 1 presents the choice proportions for each of the three response options across the four experimental outcomes. For three of the four outcomes—activated and biased, not activated and biased, and activated and not biased—the response option consistent with the logical implication of the hypothesis received the highest proportion of choices. This pattern indicates that participants tended to select the normatively appropriate response for these three outcomes. In contrast, for the not-activated and not-biased outcome, the response option corresponding to the logical contrapositive did not receive the highest proportion of choices.

To examine these patterns statistically, chi-square tests were conducted separately for each of the four outcomes to test whether the distributions of responses deviated from a uniform distribution. The results revealed significant differences among the three response options for the activated and biased outcome, $\chi^2(2) = 274.35$, $p < .001$; the not activated and biased outcome, $\chi^2(2) = 6.31$, $p < .05$; and the activated and not biased outcome, $\chi^2(2) = 61.73$, $p < .001$. In contrast, the test for the not-activated and not-biased outcome was not significant, $\chi^2(2) = 5.38$, $p = .06$. Thus, participants' responses were not reliably differentiated across the three options for the contrapositive case.

These results indicate that participants systematically differentiated among response options for three outcomes but failed to do so for the contrapositive outcome. In other words, participants did not consistently treat the contrapositive case as confirmatory evidence for the hypothesis, even though their responses to the other outcomes were broadly consistent with the logical structure of the hypothesis.

The analyses above were based on aggregated response distributions, leaving open the possibility that the observed pattern reflects heterogeneity in individual reasoning strategies. To address this issue, latent class analyses were conducted on responses to the four outcomes. Models with two, three, and four latent classes were estimated. The Bayesian Information Criterion (BIC) values were 2253.74 for the two-class solution, 2278.50 for the three-class solution, and 2305.68 for the four-class solution. Because the two-class model yielded the lowest BIC, it was selected as the best-fitting solution.

Based on the conditional response probabilities, the two classes were interpreted as follows. Class 1, labeled the *average class*, exhibited a response pattern closely resembling the aggregate results. The estimated population proportion for this class was .55, indicating that the overall pattern was not merely an artifact of averaging across qualitatively distinct participants.

Class 2 was labeled the *satisficing class*, as its response pattern suggested a focus on whether the antecedent (A) and the consequent (B) of the conditional statement ("if A, then B") were satisfied, rather than on the logical validity of the inference. Specifically, the conditional probabilities for choosing "it confirms the hypothesis" were relatively high when the antecedent was satisfied and lowest when both the antecedent and the consequent were unsatisfied. Conversely, the probabilities for choosing "it disconfirms the hypothesis" were relatively high when the antecedent was unsatisfied and highest when both the antecedent and the consequent were unsatisfied. This pattern suggests that participants in this class tended to equate confirmation with the satisfaction of the antecedent and the consequent.

Importantly, the conditional response probabilities in neither class conformed to the predictions of formal logic. Participants in Class 1 did not reliably treat the contrapositive as confirmatory evidence, whereas participants in Class 2 appeared to approach the task as an evaluation of premise and outcome satisfaction rather than as a problem of logical inference. In both classes, the probability of judging the contrapositive outcome as confirming the hypothesis was the lowest among the three response options. Taken together, the latent class analysis indicates that participants did not engage in hypothesis evaluation in a manner consistent with formal logical norms and did not regard contrapositive evidence as supportive of the hypothesis.

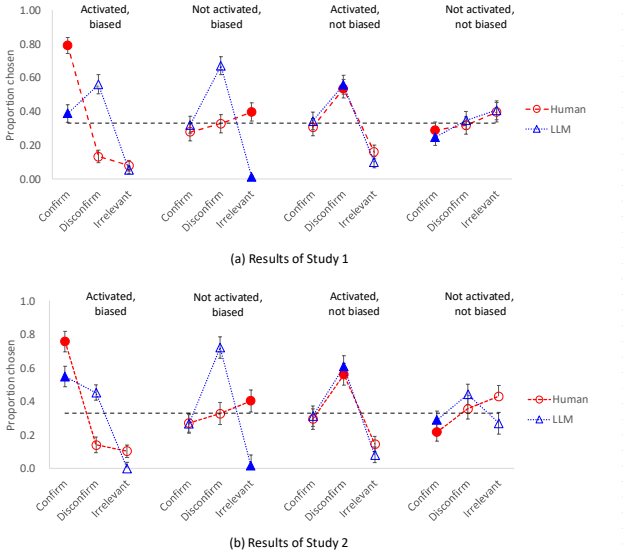

Figure 1. Results of the two studies and LLM simulation: Top and bottom panels indicate the results of Study 1 and Study 2, respectively. Filled shapes in the figures indicate the correct answers for the questions. Error bars indicate 95% confidence intervals.

Study 2

Study 2 aimed to examine whether the pattern observed in Study 1 would persist when the experimental scenario was framed as a replication of a prior study. As discussed above, the paradox of the contrapositive in psychological experiments has particularly important implications in the context of replication. In such cases, results that completely fail to reproduce the original findings may nevertheless, from a logical standpoint, be interpreted as supporting the original hypothesis.

Method

Two hundred and fourteen undergraduate students participated in Study 2 in exchange for course credit. All procedures, including stimulus presentation and data collection, were conducted online using Google Forms. The experimental procedure was nearly identical to that of Study 1, with the sole exception that the cover story framed the described psychological experiment as a replication study.

Specifically, whereas Study 1 described a situation in which a researcher proposed a hypothesis and conducted an experiment to test it, Study 2 described a situation in which a researcher conducted a replication study to examine the reliability of previously reported findings. Detailed descriptions of the experimental materials and data are available on OSF.

Table 1. Results of the two class solutions of laten class analysis.

		Study 1			Study 2		
Class1		Confirm	Disconfirm	Irrelevant	Confirm	Disconfirm	Irrelevant
	Activated, biased	0.97	0.02	0.01	0.96	0.11	0.03
	Not activated, biased	0.21	0.32	0.47	0.04	0.37	0.59
	Activated, not biased	0.00	1.00	0.00	0.00	0.99	0.01
	Not activated, not biased	0.28	0.23	0.49	0.15	0.20	0.64
Class2							
	Activated, biased	0.60	0.25	0.15	0.60	0.24	0.16
	Not activated, biased	0.34	0.34	0.32	0.46	0.29	0.25
	Activated, not biased	0.63	0.04	0.33	0.53	0.22	0.25
	Not activated, not biased	0.29	0.41	0.29	0.26	0.48	0.64

Results and discussion

Figure 2 presents the selection rates for each response option across the four experimental outcomes in Study 2. Overall, the pattern closely resembles that observed in Study 1. For three of the four outcomes, the response option consistent with the logical implication of the hypothesis received the highest proportion of selections. In contrast, for the not-activated and not-biased outcome, the option corresponding to the logical contrapositive did not receive the highest selection rate; notably, the most frequently chosen option was not "irrelevant," which would be the normatively appropriate response.

As in Study 1, chi-square tests were conducted separately for each outcome to examine whether response distributions deviated from a uniform distribution. Significant differences among the three response options were found for the activated and biased outcome, $\chi^2(2) = 173.31$, $p < .001$; the activated and not-biased outcome, $\chi^2(2) = 59.96$, $p < .001$; and the not-activated and not-biased outcome, $\chi^2(2) = 15.29$, $p < .001$. In contrast, the distribution for the not-activated and biased outcome did not significantly deviate from uniformity, $\chi^2(2) = 5.53$, $p = .06$. Although this pattern of significance was not identical to that observed in Study 1, the results again indicate that participants failed to select the normatively appropriate response in the not-activated and not-biased condition, while they were largely able to do so in the activated conditions.

Latent class analysis was conducted following the same procedure as in Study 1 and yielded a highly similar solution. The BIC values for the two-, three-, and four-class models were 1647.32, 1669.56, and 1700.18, respectively, indicating that the two-class model provided the most parsimonious and best-fitting representation of the data. Based on the conditional response probabilities, the two classes were again interpreted as an *average class* and a *satisficing class*. The average class exhibited response patterns closely aligned with the aggregate choice proportions across the four outcomes, whereas the satisficing class showed a tendency to base judgments on whether the antecedent and consequent of the conditional were satisfied, rather than on its logical structure.

In sum, despite minor differences in the detailed pattern of statistical significance, Study 2 largely replicated the key findings of Study 1. Across both studies, participants systematically failed to treat the contrapositive outcome as

confirmatory evidence for the hypothesis, even when the experimental context was explicitly framed as a replication.

Large language model simulation

The results of Studies 1 and 2 consistently demonstrated that participants did not treat the contrapositive as confirmatory evidence for the hypothesis. Moreover, a substantial subset of participants appeared to approach the task not as hypothesis testing in a logical sense, but rather as a check of whether the antecedent and the consequent of the conditional statement were satisfied. This raises an important question: Is this tendency specific to human reasoning, or can it also be reproduced by artificial intelligence systems? To address this issue, the present study conducted simulations using a large language model (LLM), applying the same experimental protocols as in Studies 1 and 2 and comparing the model's responses with the observed human data.

Specifically, we conducted a simulation study using ChatGPT-4.0 to generate surrogate participant responses. The model was instructed to assume the role of a non-expert lay participant and to respond based on intuitive judgment rather than formal logical reasoning. All responses were generated in a single batch using a fixed model version and fixed generation parameters. Each simulated participant was presented with the same description of the hypothetical psychological experiment used in Studies 1 and 2, which tested the hypothesis that activation of a mental concept can bias subsequent judgments. The simulated participants then evaluated four possible outcome patterns crossing the presence or absence of concept activation and judgment bias. For each outcome, they judged whether it confirmed the hypothesis, disconfirmed it, or was irrelevant. The order of the four outcome patterns was randomized independently for each simulated participant. All responses were generated with identical parameters across participants (temperature = 0.8).

This simulation procedure was conducted separately for the experimental protocols corresponding to Study 1 and Study 2. In each simulation, responses were generated for 300 simulated participants. To assess the stability of the resulting response patterns, these samples were treated as approximations of the underlying response distributions, and a nonparametric bootstrap procedure with 10,000 resamples of size 300 was applied. Summary statistics were computed for each resample, allowing estimation of sampling variability without issuing additional model queries.

The results of the LLM simulations, shown in Figure X, indicate that the model's responses to the two not-biased outcomes closely matched those of human participants in both Study 1 and Study 2. For the not-activated and not-biased outcome in particular, the LLM's responses were highly similar to the human data in Study 1 and differed only slightly in the proportion of "irrelevant" responses in Study 2. Taken together, these findings suggest that the LLM was generally able to approximate human judgment patterns for outcomes in which no bias was observed.

In contrast, the LLM's responses to the two biased outcomes diverged from the human data in two important respects. First, for the not-activated and biased outcome, the model's response distribution differed from that of human participants, although it was relatively consistent across Study 1 and Study 2. Second, for the activated and biased outcome, the model's responses differed from the human data and were also inconsistent between the two studies.

Overall, the LLM simulation results support three main conclusions. First, the LLM closely approximated human judgments for disconfirmatory outcomes. Second, the model failed to reproduce human judgments for outcomes that were normatively irrelevant, despite exhibiting internally consistent response patterns across studies. Third, the LLM did not successfully simulate human judgments for confirmatory outcomes, and its response patterns for these outcomes varied across experimental contexts. In sum, while the LLM captured key aspects of human reasoning for disconfirmatory evidence, it performed poorly in reproducing human judgments for confirmatory evidence.

General discussion

The main findings of the present paper can be summarized as follows. First, the results demonstrate that Hempel's paradox can arise under experimental conditions that closely resemble those of typical cognitive psychology research. Specifically, even when an experiment fails to generate—or to replicate—results that are normatively ideal for a hypothesis, those results may nevertheless be interpreted as supporting the hypothesis. Second, participants were generally unable to treat the contrapositive as confirmatory evidence, suggesting that they genuinely experienced this situation as paradoxical. Third, the large language model (LLM) failed to reproduce the human response patterns across the two studies, particularly in the condition that was most favorable for confirming the hypothesis.

To the best of our knowledge, this study provides the first empirically grounded demonstration of how Hempel's paradox can arise in a realistic scientific context. Although Hempel's paradox has long been discussed as a philosophical illustration of the limits of purely logical accounts of scientific reasoning, few studies have examined how the paradox might manifest in actual research practices. The present findings suggest that ordinary psychological experiments can give rise to situations structurally equivalent to Hempel's paradox and offer a concrete illustration of how such a paradox may emerge in practice.

In addition, the present study provides further evidence that people have difficulty treating the contrapositive as evidence for a hypothesis. A central conclusion in the literature on human reasoning is that people exhibit confirmation bias in hypothesis testing (Nickerson, 1998; Wason, 1960, 1968), often failing to regard disconfirming instances as relevant because they do not interpret the contrapositive as confirmatory. Previous studies have typically examined this tendency using tasks that involve searching for relevant evidence (Wason, 1968), evaluating

specific instances (Wason, 1960), or comparing the degree of support provided by different types of evidence (McKenzie & Mikkelsen, 2000).

In contrast to these approaches, the present study adopts a distinctive method by embedding hypothesis testing within a vignette that closely resembles a psychological experiment and by explicitly asking participants to judge whether the experimental outcomes confirm the hypothesis. In this sense, the present work may be regarded as one of the first attempts to directly examine how people evaluate experimental results themselves as confirmatory or non-confirmatory evidence.

The analyses based on the LLM simulations indicate that the model was able to approximate human judgments only in the case of disconfirmatory evidence. Recent studies have investigated whether LLMs solve reasoning and decision-making tasks in ways that resemble human cognition (Binz et al., 2023; Lampinen et al., 2024), with mixed conclusions. The present findings similarly reveal a nuanced pattern: while the LLM did not successfully reproduce human judgments across all outcome types, it generated response patterns that closely matched human judgments in the disconfirmation condition.

This result is noteworthy because much of the existing literature comparing human and LLM performance has emphasized overall accuracy or aggregate similarity, often highlighting parallels between the two. By contrast, the present findings suggest that humans and LLMs may share similar strategies for evaluating disconfirmatory evidence, while diverging substantially in how they treat confirmatory evidence. This asymmetry has important implications for the development of computational theories of confirmation and disconfirmation.

Finally, the present study raises two important directions for future research. First, additional empirical instances of Hempel's paradox in psychological science should be explored. One promising domain is neuroimaging research, where correspondence between patterns of brain activity and behavioral measures is often treated as a key criterion for empirical inference. Examining how researchers and laypeople interpret correspondence and non-correspondence in such contexts may provide further insight into the scope of Hempel's paradox.

Second, further work is needed to clarify how humans and LLMs resolve confirmation and disconfirmation at a computational level. The results of the present studies consistently suggest that neither humans nor LLMs approach hypothesis testing in a strictly logical manner, and that the strategies they employ differ in systematic ways. This raises a more fundamental question: how many distinct strategies for hypothesis testing are available in principle, and which of these strategies are actually adopted by humans and by artificial systems? Addressing this question should be a central goal of future research on human reasoning and artificial intelligence.

Acknowledgments

The authors would thank to Keiga Abe (Wako university) and Sachiko Takagi (Tokiwa university) for their help with collecting the data.

References

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology, 71*(2), 230–244. https://doi.org/10.1037/0022-3514.71.2.230

Binz, M., et al. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences, 120*(6), e2218523120. https://doi.org/10.1073/pnas.2218523120

Evans, J. St. B. T., & Lynch, J. S. (1973). Matching bias in the selection task. *British Journal of Psychology, 64*(3), 391–397. https://doi.org/10.1111/j.2044-8295.1973.tb01365.x

Hempel, C. G. (1945). Studies in the logic of confirmation (I). *Mind, 54*(213), 1–26. https://doi.org/10.1093/mind/LIV.213.1

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*(2), 263–291. https://doi.org/10.2307/1914185

Kao, S.-F., & Wasserman, E. A. (1993). Assessment of covariation by humans and animals. *Psychological Bulletin, 114*(1), 112–130. https://doi.org/10.1037/0033-2909.114.1.112

Lampinen, A. K., et al. (2024). Can language models reason about novel problems? *Cognitive Science, 48*(1), e13318. https://doi.org/10.1111/cogs.13318

Levin, I. P., Wasserman, E. A., & Kao, S.-F. (1993). Multiple methods for examining biased information use in contingency judgments. *Organizational Behavior and Human Decision Processes, 55*(2), 228–250. https://doi.org/10.1006/obhd.1993.1029

Lipe, M. G. (1990). An examination of the information used in the evaluation of conditional probabilities. *Organizational Behavior and Human Decision Processes, 46*(1), 1–21. https://doi.org/10.1016/0749-5978(90)90002-T

McKenzie, C. R. M., & Mikkelsen, L. A. (2008). A Bayesian view of covariation assessment. *Cognitive Psychology, 40*(2), 123–157. https://doi.org/10.1006/cogp.1999.0724

Mussweiler, T., & Strack, F. (1999). Hypothesis-consistent testing and semantic priming in the anchoring paradigm: A selective accessibility model. *Journal of Experimental Social Psychology, 35*(2), 136–164. https://doi.org/10.1006/jesp.1998.1364

Mussweiler, T., & Strack, F. (2000). The use of category and exemplar knowledge in the solution of anchoring tasks. *Journal of Personality and Social Psychology, 78*(6), 1038–1052. https://doi.org/10.1037/0022-3514.78.6.1038

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*(2), 175–220. https://doi.org/10.1037/1089-2680.2.2.175

Schustack, M. W., & Sternberg, R. J. (1981). Evaluation of evidence in hypothesis testing. *Psychological Bulletin, 90*(3), 559–572. https://doi.org/10.1037/0033-2909.90.3.559

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review, 90*(4), 293–315. https://doi.org/10.1037/0033-295X.90.4.293

Waldmann, M. R., & Hagmayer, Y. (2005). Seeing versus doing: Two modes of accessing causal knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(2), 216–227. https://doi.org/10.1037/0278-7393.31.2.216

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology, 12*(3), 129–140. https://doi.org/10.1080/17470216008416717

Wason, P. C. (1966). *Reasoning*. In B. Foss (Ed.), *New horizons in psychology* (pp. 135–151). Penguin.

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology, 20*(3), 273–281. https://doi.org/10.1080/14640746808400161

Wasserman, E. A., Dorner, W. W., & Kao, S.-F. (1990). Contributions of specific cell information to judgments of interevent contingency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(3), 509–521. https://doi.org/10.1037/0278-7393.16.3.509