Humans generate auxiliary hypotheses to resolve conflicts in observational data

Trisevgeni Papakonstantinou, Kuan Iao Leong & David Lagnado
University College London
Department of Experimental Psychology
26 Bedford Way, London WC1H 0APA

Abstract

Although research in the area of belief updating has flourished in the last two decades, most studies do not treat beliefs as part of a complex and interactive network. In this study, we investigate humans' use of auxiliary hypotheses as a mechanism to avoid belief updating in light of conflicting information. In Experiment 1, we replicate an unpublished study by Kahneman and Tversky, introducing two additional domain conditions (N=119). Participants construct an initial model, express a prior belief, and face conflicting information. They are then prompted to provide an explanation. Across three domains, only 37% of responses demonstrate belief updating, by attributing the information conflict to the original report being unreliable or invalid. In Experiment 2 (N=29), a within-participants manipulation of credibility shows no effect on generating auxiliary hypotheses. Even in the presence of credibility cues to explain away information conflicts by invoking the reliability of either source, participants instead generated auxiliary hypotheses to resolve them in 27% of the cases.

Keywords: Keywords: belief updating; auxiliary hypotheses; credibility; mental models; Duhem-Quine thesis

Introduction

Despite the imperative for adaptive reasoning, human minds exhibit a pronounced aversion to conceptual change, resulting in the enduring resilience of beliefs even when confronted with compelling evidence to the contrary (Kube & Rozenkrantz, 2021). While commonly depicted as a flaw in human reasoning (i.e. a "bias"), this resistance to change may, in fact, be an outcome of rational principles at play (Gershman, 2018). Beliefs, rather than existing in isolation, form complex systems that strive for coherence both within themselves and in conjunction with other beliefs (Thagard, 1989). Successful revision of incorrect models depends crucially on the willingness to re-evaluate and modify critical false beliefs in this intricate web of cognitive constructs (Chi, 2000). When confronted with anomalous information, especially information that attacks self-identifying beliefs, individuals tend to employ a number of strategies to resolve the conflict, while striving maintain the original central belief and the coherence of the model. Observations that appear to contradict a central hypothesis can be "explained away" by changing or introducing auxiliary hypotheses - and sometimes, for good reason (Gershman, 2018).

The Duhem-Quine thesis, a key concept in the philosophy of science, significantly intersects with the landscape of belief and mental model updating discussed above. Proposed by Pierre Duhem and later expounded by Willard Quine, this thesis challenges the traditional notion of isolating individual hypotheses for testing. Instead, it asserts that theories are interconnected and any observation can be protected from refutation by adjusting auxiliary hypotheses (Quine, 1951). In the context of belief revision, the Duhem-Quine thesis underscores the interdependence of beliefs within a broader system, reflecting the idea that the revision of a single belief may necessitate adjustments throughout the entire web of interconnected beliefs. This holistic perspective aligns with the observed resistance to change in human reasoning, emphasizing the complex and interwoven nature of belief systems, akin to the interrelated theories in scientific paradigms. In the pursuit of maintaining mental model coherence, individuals could adeptly "explain away" apparent inconsistencies by introducing or adjusting auxiliary hypotheses, a process with deep roots in maintaining stability and coherence within our cognitive frameworks (Thagard, 1989).

Various factors influence this process. Evidence suggests that humans tend to generate auxiliary hypotheses to resolve conflicts between observational data and prior beliefs when they are engaged in processes such as diagnostic hypothesis generation (Thomas, Dougherty, Sprenger, & Harbison, 2008). Individuals are less likely to consider alternative hypotheses if they already have one that fits the data, an effect known as "satisficing" (Garst, Kerr, Harris, & Sheppard, 2002). Computational modeling studies suggest that inferences sometimes deviate from rationality due to the self-generation of hypotheses; people are likely to believe hypotheses that are self-generated are less likely to be true compared to those generated by others and presented to them (Dasgupta, Schulz, & Gershman, 2017). The order the information is presented is significant. Since their seminal 1974 study, a substantial body of literature has demonstrated the effect that Tversky and Kahneman first coined as the "anchoring heuristic" (Tversky & Kahneman, 1978), a strong bias to limit the deviation from an initially presented hypothesis (for a theoretical review see Lieder, Griffiths, Huys, & Goodman, 2017). Uncertainty mediates this effect; order effects in the process of human belief revision become apparent when there is lack of confidence in the initial hypothesis, and seem to subside where confidence is high (Wang, Zhang, & Johnson, 2000).

Credibility seems to also play a central role. Research on consumer decision making indicates that people are more

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

likely to trust information from sources they know personally, and that the level of information credibility affects their purchase decisions (Cooley & Parks-Yancy, 2019). When it comes to learning from information presented by social partners, the extent to which individuals adjust their beliefs based on that information is significantly predicted by factors such as the perceived competence, reliability, and the level of trust attributed to these partners (Pilditch, Madsen, & Custers, 2020). Credibility cues are used to guide judgements of trustworthiness and subsequent decision-making, with evidence suggesting a preference for credibility over cognitive or operational utility (Gugerty & Link, 2020).

The cognitive function of ad hoc auxiliary hypotheses generation and the process of reasoning out of incoherence, has been the object of much theorising (Mandelbaum, 2018; Johnson-Laird, Girotto, & Legrenzi, 2004). However, limited empirical research has directly explored the role of auxiliary hypotheses in belief revision, and little is known about what influences individuals' approach to integration of auxiliary hypotheses in their mental models. Synthesising previous research on belief updating and social learning, here we aim to explore factors that affect the likelihood of auxiliary hypotheses generation during an information conflict. In Experiment 1 we replicate a classic paradigm by Kahneman and Tversky, aiming to elicit participants' explanations for a series of information conflicts in different domains. We hypothesised that the results from Kahneman & Tversky's original study would be replicated. They found that less than 20% of participants invoked the invalidity of the original information as an explanation (Kahneman & Tversky, 1982). Additionally, we predicted that confidence will play a role; likelihood of updating will be lower when ratings of confidence in the initial hypothesis are higher (i.e. stronger priors). Finally, we explore the effect of information domain on likelihood of information integration - comparing "soft" and "hard" evidence, as represented by scenarios base in the social and physical domains respectively. In Experiment 2 we aimed to investigate how different combinations of information credibility influence whether auxiliary hypotheses are being used or not for the integration of disconfirming evidence. We hypothesised that when the credibility level of prior and posterior information is similar, there will be higher likelihood of participants generating auxiliary hypotheses, compared to cases where there is an imbalance of credibility.

Experiment 1

Methods

Preregistration This study's data collection procedure, experimental design, materials and measures, as well as the main hypotheses were registered on the Open Science Framework (`https://doi.org/10.17605/OSF.IO/MKSWF`).

Participants and design We recruited 121 participants via Prolific (M_{age} = 40.13, SD_{age} = 12.00, N_{female} = 102). No participants failed the attention checks. Two participants demonstrated lack of understanding in their response to the open-ended question, so they were excluded in the counts for this condition only, and the overall analysis. The experiment has a quasi-experimental design, where the domain of the information was manipulated within participants, with three levels: 'personality', 'social', 'physical'. The domain manipulation acted mainly as a control with an aim to inform the generalisability of the findings beyond the single domain investigated by Khaneman and Tversky, especially as a difference in domain often implies a difference in priors.

Materials and procedure

All participants completed all three domain conditions, with the order randomised and counterbalanced across participants. As this study is a replication and extension of the Tom W. study by Kahneman & Tversky (1982), we used their original materials for the 'personality' condition and added two more conditions, 'social' and 'physical' to investigate whether the domain of the vignette might have an effect, replicating the same vignette structure. We diverged from their original design by adding confidence ratings; participants were prompted to indicate their level of confidence in their prior and posterior judgements. The procedure was the same for all three conditions; First, participants were presented with a short report, which they were explicitly told was produced by a source of uncertain validity.

> **Vignette 'Personality / Replication'** *The following is a personality sketch of Tom W written during Tom's senior year in high school by a psychologist, on the basis of psychological tests of uncertain validity: Tom W. is of high intelligence, although lacking in true creativity. He has a need for order and clarity, and for neat and tidy systems in which every detail finds its appropriate place. His writing is rather dull and mechanical, occasionally enlivened by somewhat corny puns and by flashes of imagination of the sci-fi type. He has a strong drive for competence. He seems to have little feel and little sympathy for other people and does not enjoy interacting with others. Self-centered, he nonetheless has a deep moral sense.*

After reading the first part of the vignette, participants were asked to make a prediction based on the information in the original report. They are prompted to rank nine fields of graduate specialization in order of the likelihood that Tom W. is now a graduate student in each of these fields (e.g. "Computer science", "Medicine", "Humanities and education"). They were also asked to indicate their confidence in their prediction on a slider scale. This first stage was followed by a brief attention check. The second part of the vignette reveals information about the true turn of events, which is in conflict with what the original report had led them to assume.

> **Vignette 'Personality / Replication'** *In fact, Tom W.*

is a graduate student in the School of Education and he is enrolled in a special program of training for the education of handicapped children.

Finally, participants were asked to briefly outline the theory which explains the relationship between the information in the original report, and what they learned was the true turn of events. They wrote their answer into an open textbox with unlimited character length. This open-ended question allowed us to determine whether participants were able to revise their prior models in favour of the new conflicting information. Materials for the 'social' and 'physical' conditions are available on OSF. Lastly, they were asked to indicate their confidence in their response, similarly as before. At the end of the experiment participants provided demographic information.

Results

Qualitative analysis Based on participants' free text responses, we developed a coding framework for theories generated to explain the conflict between the original report and follow-up information. The coding framework had three main categories. Participants' responses were coded by the first author and a research assistant.

The first category, "Updating", covered all responses that invoked the invalidity of the original report as an explanation for the conflict, reflecting successful revision of the original model and integration of the new information. Responses in that category were coded as such whether they invoked the uncertainty probe that was provided, or some other reason why the report is not valid. Nine responses (8%) fell under this category in the 'personality' condition, 51 responses (42%) in the 'social' condition, and 73 responses (60%) in the 'physical' condition.

The second category, "Auxiliary hypothesis", comprised responses invoking auxiliary hypotheses to explain the relationship between the original and conflicting information. Responses in this category were coded as such in cases where participants either generated new information (nodes in the model), or new links between information they were given, in order to explain the relationship. To demonstate what is referred to as an "auxiliary hypothesis" two representative quotes follow:

"I feel that Tom may have had some trouble growing up and wants to help other children"

"The CEO and the dismissive female colleague are in a clandestine relationship"

89 responses (75%) fell under this category in the 'personality' condition, 57 responses (47%) in the 'social' condition, and 42 responses (35%) in the 'physical' condition.

The final category, "Failure to integrate", covered responses that demonstrated failure to integrate the two pieces of information. Responses in this category expressed surprise and the feeling that the outcome revealed in the second stage of the vignette did not make sense. Eight responses (7%) fell under this category in the 'personality' condition, 15 responses (12%) in the 'social' condition, and six responses (5%) in the 'physical' condition.

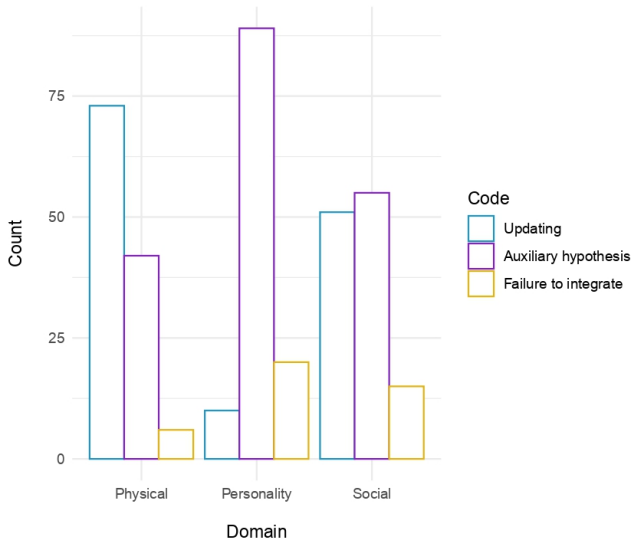

Figure 1: Prevalence of explanatory strategies in each domain

In all three conditions, the majority of participants ($> 63\%$ across all conditions ranked the 'true' outcome in the three last positions (7th, 8th or 9th) in terms of its likelihood when asked to make a prediction. Notably, a considerable percentage of participants (9%) ranked the 'true' outcome as the most likely in the 'physical' condition.

Regression analysis We implemented a Bayesian approach, which was not specified in the pre-registered analysis plan (see `https://osf.io/fkj27/`), though we did not deviate from the specified model. We fitted Bayesian mixed-effect logistic regression models with pseudo-Bayesian model averaging and a Bayesian bootstrap (Hinne, Gronau, van den Bergh, & Wagenmakers, 2020; Yao, Vehtari, Simpson, & Gelman, 2018; Bürkner, 2018). We specified three models incrementally adding predictors based on our degree of confidence in their effects according to previous findings. Participant ID was included in all models as a random effect. Our starting model included only the prior ranking of the true outcome with a global regularizing prior $\mathcal{N}(0, 0.50)$ set for all (standardized) predictors. We then added domain along with two-way interactions, and finally difference in confidence rating from prior to posterior, again including all two-way interactions. At each step, we computed Pseudo-BMA weights with Bayesian bootstrap to assign weights to each model, then sampled from the averaged posterior distribution to compute the central tendency and uncertainty of predictor effects.

Prior ranking of the true outcome did not have a significant effect on likelihood of belief updating. Domain had a

Table 1: Bayesian mixed-effects logistic regression with model averaging and regularizing priors predicting belief updating

		95% CI	
Parameter	Estimate	Lower	Upper
Intercept	−0.84	−1.16	−0.54
Prior	−0.19	−0.57	0.14
Social (Personality)	0.49	0.13	0.84
Physical (Personality)	1.23	0.87	1.62
$\Delta Confidence$	0.45	0.16	0.77
Prior : Social (Personality)	0.20	−0.21	0.62
Prior : Physical (Personality)	0.69	0.29	1.11
Prior : $\Delta Confidence$	−0.04	−0.31	0.24
Social (Personality) : $\Delta Confidence$	0.01	−0.37	0.40
Physical (Personality) : $\Delta Confidence$	−0.39	−0.75	−0.02

significant effect, with both the social (0.49, 95% CI: [0.13, 0.84]) and physical (1.23, 95% CI: [0.87, 1.62]) domains being associated with a higher predicted likelihood of updating, compared to the original personality domain. A larger difference between prior and posterior confidence (with posterior being higher) was also associated with higher likelihood of belief updating (0.45, 95% CI: [0.16, 0.77]). There was significant interaction between the prior judgement and domain, with higher ranking of likelihood of the true outcome being associated with higher likelihood of updating for the physical domain, compared to the personality domain (0.69, 95% CI: [0.29, 1.11]). Finally, a smaller but significant interaction effect was found between domain and difference in confidence; with larger difference in confidence being associated with lower likelihood of updating for the physical domain, compared to the personality domain (-0.39, 95% CI: [-0.75, -0.02]).

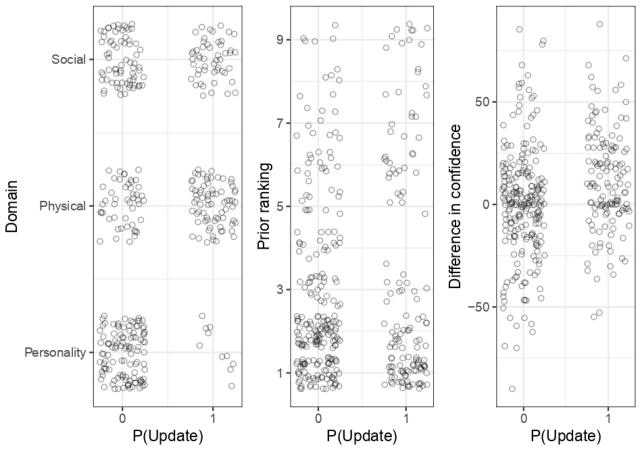

Figure 2: Updating probability by domain, prior ranking of true outcome and difference in confidence. Raw data has been jittered for readability.

Overall, these findings highlight the generation of auxiliary hypotheses as a prominent strategy in cases of information conflict. In the original domain as well as the added 'social' condition, the majority of participants did not express any doubt about the validity of the original report in response to disconfirming information. In total, only 133 of the responses (37%) across all domains invoked the invalidity of the original report as the explanation for the conflict, reflecting belief updating. These results support our main hypothesis and are in line with findings from Kahneman and Tversky's original study. However, when taken individually, this hypothesis is only supported for the 'personality' and 'social' domains, but not the 'physical' one. The results suggest the use of auxiliary hypotheses as a prominent alternative avenue for the explanation of the conflict, represented by more than 35% of responses in all conditions, with an average prevalence of 52% across all responses.

Experiment 2

Methods

In Experiment 2, we introduced a credibility manipulation by embedding cues regarding the source and quality of the presented information within the vignettes. This feature aimed to provide participants with an obvious avenue for resolving cognitive conflict, particularly in cases of imbalance (high credibility source versus low credibility source). It was hypothesized that when participants are asked to explain the conflict and provide their opinion on the ground truth, they would invoke the low credibility of one or both of the sources to invalidate the corresponding information. It ensured that if participants resorted to auxiliary hypotheses to resolve the conflict, it was not for lack of easier options to explain it away. This would suggest the presence of a strong cognitive bias favouring the use of auxiliary hypotheses to maintain model coherence, rather than disconfirmation.

Preregistration The data collection procedure, design, materials and measures, as well as the main hypotheses and analysis plan were pre-registered on the Open Science Framework (`https://doi.org/10.17605/OSF.IO/WATG5`).

Participants and design We recruited 29 participants via Prolific (M_{age} = 34.55, SD_{age} = 10.20, N_{female} = 21). No participants failed the attention checks. Sample size calculations were based on a power analysis to detect an effect size with critical alpha of 0.05 with 80% power, which set the minimum required sample size at N = 24. All participants passed the attention checks. We manipulated information credibility within participants, with 4 levels: high prior-low posterior (i.e., prior information with high credibility and posterior information with low credibility), low prior-high posterior, low prior-low posterior, high prior-high posterior. Domain was randomised across conditions and between participants.

Materials and procedure

A similar task structure to Experiment 1 was used. Participants were first presented with a short report, this time with a

named source and credibility cues. There were four different vignettes in four separate domains: 'Political Election', 'Criminal Trial', 'Business Plans', and 'Scientific Finding'. These all represented new scenarios and do not map on to any used in Experiment 1. As in the previous Experiment, different domains were employed to inform generalisability, rather than explore specific domain effects. Materials for all conditions are available on OSF. An example of the first condition (high - low credibility) is presented below:

> **Vignette 'Political Election'** *During a political election, an anonymous opinion piece surfaces on a political blog. The author raises concerns about a candidate's trustworthiness. In a detailed exposé, he mentions a high-profile fundraising event held last year, where the candidate was seen socializing with individuals implicated in financial improprieties and corruption scandals. The piece further suggests potential ethical concerns by referencing accounts from former staff members and anonymous insiders who claim the candidate was privy to, and possibly complicit in, embezzlement schemes and bribery attempts within their previous professional circles.*

After reading the first part of the vignette, participants were asked to make a judgement on the subject matter based on the information in the original report, by providing a rating on a slider scale representing, in this case, their perception of the candidate's character(e.g. 0 being completely unethical, 100 being exemplar). They were also asked to indicate their confidence in their judgement on a slider scale. The second part of the vignette reveals additional information on the subject matter, this time from a different source, which is in conflict with what the original report had led them to believe.

> **Vignette 'Political Election'** *Subsequently, a respected news outlet releases an investigative report detailing the qualifications and accomplishments of the same candidate, emphasizing transparency and ethical governance. This news outlet is known for its well-researched and impartial accounts and for equal treatment of different political parties. In a comprehensive analysis, the report provides extensive evidence refuting the claims made in the anonymous opinion piece. It includes verified testimonials from reputable sources present at the event, indicating that the candidate engaged in ethical discourse and did not associate with individuals involved in any controversial activities. The investigative report meticulously examines the candidate's professional background, highlighting achievements, and showcasing endorsements from colleagues and community leaders attesting to their integrity. The news outlet also presents a timeline of the candidate's career, demonstrating a consistent commitment to ethical practices.*

In the final section, participants were asked to briefly outline their interpretation of the incident based on the ground truth and explain the discrepancy between the two accounts in an open text box with unlimited character length. Next, they were asked to indicate their judgement on the subject matter again, similarly as before and to provide a confidence rating on that. Finally, they were asked to rate the credibility of the two sources. At the end of the experiment participants provided demographic information.

Results

We followed the same approach and coding framework as Experiment 1. In addition to the three original categories which remained consistent, a fourth code was created, "Discounting", to reflect cases where participants discounted the second piece of information as invalid or inconsequential. Responses in this category were coded as such in cases where participants did integrate both accounts, but invoked the invalidity or negligibility of the second one in favour of the first. Four responses across all conditions were excluded as invalid or off-topic.

In the *low - high credibility* condition, 10 responses (35%) were coded as "Updating", invoking invalidity of the original report in favour of the second, 11 (38%) were coded as invoking "Auxiliary hypotheses", two (7%) as "Discounting" the follow-up report, and six (21%) were coded as "Failure to integrate". In the *high - low credibility* condition, five responses (17%) were coded as reflecting "Updating", seven (24%) as invoking "Auxiliary hypotheses", eight (28%) as "Discounting" the follow-up report, and seven (24%) were coded as "Failure to integrate". In the *low - low credibility* condition, seven responses (24%) were coded as reflecting "Updating", six (21%) as invoking "Auxiliary hypotheses", seven (24%) as "Discounting" the follow-up report, and eight (28%) were coded as "Failure to integrate". Finally, in the *high - high credibility* condition, six responses (21%) were coded as reflecting "Updating", seven (24%) as invoking "Auxiliary hypotheses", four (14%) as "Discounting" the follow-up report, and 11 (38%) were coded as "Failure to integrate". Figure 3 presents a breakdown of the coded explanatory strategies by credibility condition.

Despite the availability of credibility cues that could easily be employed to dismiss information conflicts, the prevalence of auxiliary hypotheses was still high. Across conditions participants generated auxiliary hypotheses to resolve conflicts in 27% of the cases. Notably, that percentage is highest for the low-high credibility condition, where both order and imbalance in credibility should favour straightforward belief updating. Failure to integrate was unsuprisingly the highest in the high-high credibility conditions, where both accounts were equally convicing.

A Bayesian mixed-effects logistic regression with participant ID as a random effect showed no effect of condition on

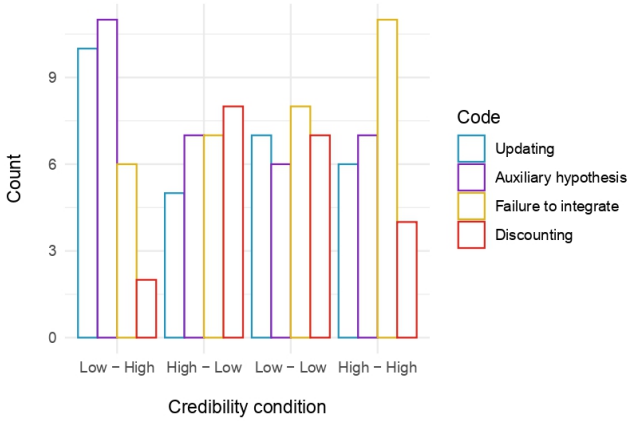

Figure 3: Prevalence of explanatory strategies by credibility condition

Table 2: Bayesian mixed-effects logistic regression with regularizing priors predicting the generation of auxiliary hypotheses

		95% CI	
Parameter	Estimate	Lower	Upper
Intercept	−1.24	−2.08	−0.56
Low-High (Low-Low + High-High)	0.38	−0.16	0.94
High-Low (Low-Low + High-High)	−0.07	−0.64	0.48
Low-Low (High-High)	−0.09	−0.68	0.49

likelihood of generating auxiliary hypotheses. The results are summarized in Table 1.

Discussion

The findings from Experiment 1 provide valuable insights into the dynamics of belief revision and the generation of auxiliary hypotheses in the face of conflicting information. Replicating Kahneman and Tversky's classic paradigm, our results support the notion that individuals often resist invoking the invalidity of the original report as an explanation for conflict, with less than 20% doing so across all domains. This finding aligns with the broader literature on resistance to belief updating.

Both the social and physical domains were associated with a higher likelihood of updating compared to the original personality domain condition. This could be attributed to the more tangible and concrete nature of evidence in these domains, making it more compelling for individuals to reevaluate their initial beliefs. Domain effects on belief revision are well documented (Müller-Otto, Garcia-Retamero, Galesic, & López, 2013). The interaction effects further illuminate the nuanced interplay of variables influencing belief revision. A larger difference between prior and posterior confidence was associated with a higher likelihood of belief updating, emphasizing the importance of uncertainty in the revision process. This supports previous findings on the role of confidence in updating (Wang et al., 2000). Moreover, the interaction between prior judgment and domain suggests that priors differentially influence the likelihood of updating in different domains. This implies that individuals may be more inclined to revise beliefs in domains where the true outcome is initially perceived as more likely. Notably, explanations for the conflict in the 'physical' condition were the least likely to invoke auxiliary hypotheses and the most likely to reflect belief updating. One interpretation of this phenomenon relates to the hypothesised cognitive function of ad hoc auxiliary hypotheses; that of a "protective belt" (Lakatos, 1970) or "psychological immune system" (Mandelbaum, 2018) emerging to protect the core beliefs that constitute our self-identity when facing a threat. Beliefs or propositions about the physical world are less likely to be coded as relevant to one's self-conception.

In the face of credibility cues designed to address information conflicts, participants resorted to generating auxiliary hypotheses to resolve such conflicts in one-third of cases. The percentage is notably lower than that in Experiment 1. This could be attributed to several factors, including the design discrepancies of the materials used. In Experiment 2, participants were presented with two comprehensive reports detailing contrasting versions of the events in question. Conversely, in Experiment 1, they only received a single report intended to induce a misleading impression, followed by short supplementary information revealing the ground truth. While credibility cues are known to significantly influence social learning and information trust, our findings suggest that the generation of auxiliary hypotheses remains a prevalent strategy even when credibility signals are present, and that there is no variation in its prevalence associated with different levels of credibility. Despite external cues suggesting the credibility of information sources, individuals appear to engage in a cognitive reflex to construct additional explanations or adjust existing mental models. Further exploration into the specific content and nature of these auxiliary hypotheses could shed light on the cognitive strategies individuals employ during belief revision. Understanding the factors that contribute to the persistence of auxiliary hypothesis generation, especially in the context of credibility cues, may provide valuable insights into the complexities of human reasoning and decision-making.

The exploratory nature of this study inherently involves a trade-off with experimental control. Further research that allows for more controlled and robust measurement of the outcome is needed. A qualitative-first approach allowed for a more nuanced exploration of the complex and multifaceted processes involved in cognitive responses to conflicting information. However, without quantitative measures designed to capture the generation of auxiliary hypotheses, the results can be interpreted as more descriptive in nature. Future studies should explore the implementation of alternative paradigms that take into account the plurality and interconnectedness of beliefs, as well as the intricate interplay between cognitive constructs and external data.

References

Bürkner, P.-C. (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. *The R Journal*, *10*(1), 395–411. Retrieved from https://doi.org/10.32614/RJ-2018-017 doi: 10.32614/RJ-2018-017

Chi, M. T. H. (2000). Self-explaining: The dual processes of generating inference and repairing mental models. In R. Glaser (Ed.), *Advances in instructional psychology: Educational design and cognitive science* (Vol. 5, p. 161-238). Lawrence Erlbaum Associates Publishers.

Cooley, D., & Parks-Yancy, R. (2019). The effect of social media on perceived information credibility and decision making. *Journal of Internet Commerce*, *18*(3), 249-269. Retrieved from https://doi.org/10.1080/15332861.2019.1595362 doi: 10.1080/15332861.2019.1595362

Dasgupta, I., Schulz, E., & Gershman, S. J. (2017). Where do hypotheses come from? *Cognitive Psychology*, *96*, 1-25. Retrieved from https://www.sciencedirect.com/science/article/pii/S0010028516302766 doi: https://doi.org/10.1016/j.cogpsych.2017.05.001

Garst, J., Kerr, N., Harris, S., & Sheppard, L. (2002, 02). Satisficing in hypothesis generation. *The American journal of psychology*, *115*, 475-500. doi: 10.2307/1423524

Gershman, S. (2018, 05). How to never be wrong. *Psychonomic Bulletin Review*, *26*. doi: 10.3758/s13423-018-1488-8

Gugerty, L., & Link, D. (2020, 07). How heuristic credibility cues affect credibility judgments and decisions. *Journal of Experimental Psychology: Applied*, *26*. doi: 10.1037/xap0000279

Hinne, M., Gronau, Q. F., van den Bergh, D., & Wagenmakers, E.-J. (2020). A conceptual introduction to bayesian model averaging. *Advances in Methods and Practices in Psychological Science*, *3*(2), 200-215. Retrieved from https://doi.org/10.1177/2515245919898657 doi: 10.1177/2515245919898657

Johnson-Laird, P. N., Girotto, V., & Legrenzi, P. (2004). Reasoning from inconsistency to consistency. *Psychological Review*, *111*(3), 640–661. doi: 10.1037/0033-295x.111.3.640

Kahneman, D., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge University Press.

Kube, T., & Rozenkrantz, L. (2021). When beliefs face reality: An integrative review of belief updating in mental health and illness. *Perspectives on Psychological Science*, *16*(2), 247-274. Retrieved from https://doi.org/10.1177/1745691620931496 (PMID: 32818386) doi: 10.1177/1745691620931496

Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of knowledge* (pp. 91–196). Cambridge University Press.

Lieder, F., Griffiths, T., Huys, Q., & Goodman, N. (2017, 06). The anchoring bias reflects rational use of cognitive resources. *Psychonomic Bulletin Review*, *25*. doi: 10.3758/s13423-017-1286-8

Mandelbaum, E. (2018). Troubles with bayesianism: An introduction to the psychological immune system. *Mind and Language*, *34*(2), 141–157. doi: 10.1111/mila.12205

Müller-Otto, S., Garcia-Retamero, R., Galesic, M., & López, A. (2013, 09). The impact of domain-specific beliefs on decisions and causal judgments. *Acta psychologica*, *144*, 472-480. doi: 10.1016/j.actpsy.2013.08.004

Pilditch, T. D., Madsen, J. K., & Custers, R. (2020). False prophets and cassandra's curse: The role of credibility in belief updating. *Acta Psychologica*, *202*, 102956. Retrieved from https://www.sciencedirect.com/science/article/pii/S0001691819300952 doi: https://doi.org/10.1016/j.actpsy.2019.102956

Quine, W. V. O. (1951). Two dogmas of empiricism. *Philosophical Review*, *60*(1), 20–43. doi: 10.2307/2266637

Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences*, *12*(3), 435–467. doi: 10.1017/S0140525X00057046

Thomas, R. P., Dougherty, M. R., Sprenger, A. M., & Harbison, J. I. (2008). Diagnostic hypothesis generation and human judgment. *Psychological review*, *115 1*, 155-85. Retrieved from https://api.semanticscholar.org/CorpusID:5651538

Tversky, A., & Kahneman, D. (1978). Judgment under uncertainty: Heuristics and biases. In P. DIAMOND & M. ROTHSCHILD (Eds.), *Uncertainty in economics* (p. 17-34). Academic Press. Retrieved from https://www.sciencedirect.com/science/article/pii/B9780122148507500085 doi: https://doi.org/10.1016/B978-0-12-214850-7.50008-5

Wang, H., Zhang, J., & Johnson, T. R. (2000). Human belief revision and the order effect. In *Proceedings of the 22nd annual meeting of the cognitive science society.* Retrieved from https://escholarship.org/uc/item/3wb4r7kf

Yao, Y., Vehtari, A., Simpson, D., & Gelman, A. (2018). Using Stacking to Average Bayesian Predictive Distributions (with Discussion). *Bayesian Analysis*, *13*(3), 917 – 1007. Retrieved from https://doi.org/10.1214/17-BA1091 doi: 10.1214/17-BA1091