The GRASP Model of Belief Updating: The Impact of Source- and Argument Conditional Expectations

Jens Koed Madsen
(j.madsen2@lse.ac.uk)
Department of Psychological and Behavioural Science, London School of Economics and Political Science, Houghton Street, WC2A 2AE, London, UK.

Abstract

Bayesian models gauge and potentially describe how people integrate source reliability and argument strength when updating beliefs, but no existing account captures their joint dynamics. We present GRASP, a process-level model that combines Bayesian-inspired weighting of source reliability with non-Bayesian extensions (a confidence-based gate and a faint-praise function), organised around five core mechanisms: confidence-based gating, source-conditional expectations, faint praise inference, reliability weighting, and bidirectional reliability updating. Our main experimental findings (N= 591) are: Stronger than expected arguments increased source reliability and moved beliefs toward speaker positions, which contrasted with weaker than expected arguments; argument quality dominated initial credentials in reliability judgments; source-conditional expectations govern reliability judgments more than belief updating. The model demonstrates how multiple phenomena emerge from one principle: evidence is evaluated relative to source- and argument conditional expectations.

Keywords: belief revision; Bayesian models; source credibility; faint praise; pragmatic inference

Introduction

Most of what we know builds on information we get from other people (Hahn et al., 2012; Mercier, 2020). When encountering reports, we face an epistemic challenge: Should I update my beliefs given what was said and who said it? This requires cognitive mechanisms for epistemic vigilance (Sperber et al., 2010). This question lies at the heart of belief revision research and has implications for political discourse, science communication, and everyday decision-making.

Among theoretical approaches, probability theory has been invoked to model this challenge (Bovens & Hartmann, 2003; Hahn & Oaksford, 2007). Bayesian models represent beliefs as subjective probabilities that update when new evidence is encountered according to normative principles, derived from Bayes' theorem (Oaksford & Chater, 2007). Crucially, Bayesian models provide precise, quantitative predictions that can be empirically tested, reducing researcher degrees of freedom and making results informative (Guest & Martin, 2021). The models have accounted for a range of phenomena, including source reliability weighting (Birnbaum & Mellers, 1983; Harris et al., 2015), argument evaluation (Hahn & Oaksford, 2007), and conflicting testimonies (Bovens & Hartmann, 2003). Moreover, they have been applied to topics like issue polarisation (Young et al., 2025), political belief revision (Coppock, 2022; Young & de Wit, 2025), discriminatory reasoning (Madsen, 2019), argument fallacies (Corner et al., 2011), and dependencies (Pilditch et al., 2025).

Across multiple studies, Bayesian approaches have been instrumental in understanding core reasoning functions related to belief updating. Belief updating depends on the reasoner's attitude strength. The argument is that confidently held beliefs are more resistant to change and have a narrower zone for finding arguments plausible. The literature reports four established observations and models that have so far not been integrated.

First, reliable sources are more persuasive than unreliable sources (Hahn, Harris, & Corner, 2009; Hahn et al., 2012). Bayesian Models integrating expertise and trustworthiness support this intuition (Harris et al., 2015; Madsen et al., 2016). Second, reliability can lead to bidirectional updating where poor arguments damage a source's reliability while strong arguments enhance it (Hahn et al., 2009). Arguments are "proof in action" of competence. When sources provide poor quality or unbelievable statements, their reliability may suffer (Bovens & Hartmann, 2003; Madsen et al., 2020).

Third, argument strength influences update magnitude, as strong evidence moves beliefs more than weak evidence (Hahn & Oaksford, 2007). The quality of reasons provided matters, not just who provides them. If evidence is strongly diagnostic of a hypothesis, recipients should update more in line with the evidence. Finally, "damned by faint praise" shows that weak positive arguments may decrease belief in the hypothesis they support (Harris, Corner, & Hahn, 2013). A job application mentioning only good fashion sense implies the absence of job-relevant strengths. As such, weaker-than-expected evidence becomes negative evidence.

Each phenomenon has received attention individually, yet no model formally integrates them despite their interactions. Faint praise depends on source reliability (Harris et al., 2013), and argument quality feeds back onto reliability judgments, which implies that belief and source evaluation are coupled processes requiring joint modelling. Fully Bayesian frameworks for source-conditional updating exist (Olsson, 2013; Merdes, von Sydow, & Hahn, 2021), but they have not directly addressed attitude-strength gating or pragmatic faint-praise inference; GRASP is therefore best read as a process-level account with Bayesian-inspired components rather than a strict Bayesian network.

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

This paper introduces the Gated Reliability and Argument Strength Processing (GRASP) model. It integrates the core reasoning aspects from the Bayesian persuasion literature into a unified framework. The GRASP model outlines five core functions that make explicit, quantitative predictions about how people update both their beliefs about hypotheses according to their assessments of source reliability and argument strength. We first present each function with its theoretical motivation and formal specification. Then, using a pilot study to set stimulus parameters, Study 1 (N = 591) tests nine directional predictions derived from the model.

The GRASP model

The GRASP model examines how a reasoner should update the belief in a hypothesis (H) when presented with a report (Rep) from a source that advocates for a position (SP) between 0 and 1. That is, rather than a speaker stating that a hypothesis is true or false (see e.g., Harris et al., 2015), the speaker can express varying degrees of belief in the hypothesis.The model incorporates source reliability (R), argument strength (AS, a perceived scalar in [0,1] indexing evidential quality), attitude strength, and perceived plausibility that interact in the updating process.It contains five core functions that each address a distinct aspect of the belief revision process. In the following, we present each function, its theoretical motivation, formal specification, and connection to the broader literature.

Function 1: Confidence-based Gating Beliefs that are confidently held resist change. Petty and Krosnick (1995) documented that attitude strength predicts resistance to counter-attitudinal messages. Tormala and Petty (2004) showed that confidence moderates belief revision resistance, with highly certain individuals showing greater resistance even to objectively strong arguments. The effect has been demonstrated across diverse domains, including political attitudes, consumer preferences, and scientific beliefs (Visser & Mirabile, 2004). Also, high confidence leads to higher processing of evidence that confirms prior beliefs while disconfirming evidence is neglected (Rollwage et al., 2020)

In the GRASP model, when a claim falls far outside what a confident reasoner considers plausible, the epistemic 'gate' closes, blocking or attenuating belief updating. This captures resistance to counter-attitudinal information among confident individuals while allowing uncertain reasoners to remain open to a wider range of claims. The gate is not binary but gradient, reflecting that plausibility judgments themselves exist on a continuum.

The standardised distance z measures how many standard deviations the speaker's position lies from the reasoner's prior belief:

$$z = \frac{|SP - P(H)|}{\sigma}$$

where σ represents uncertainty (higher values indicate greater uncertainty, hence a broader plausibility zone). The gate function uses a sigmoid transformation:

$$Gate = \frac{1}{1 + \exp(k \times (z - z_{threshold}))}$$

When $z < z_{threshold}$ (claim falls within the plausible range), Gate approximates 1 and updating proceeds normally. When $z > z_{threshold}$ (claim is see as implausible), Gate approaches 0, substantially reducing belief updating. The parameter k controls the steepness of the transition between open and closed states. This sigmoid form captures the intuition that plausibility judgments are not sharp thresholds but have a region of uncertainty. Further, we introduce a Gate Modifier where in-zone claims receive full credit/blame for argument quality while out-zone claims receive reduced credit for strong arguments, but weak arguments still incur penalties

$$Gate_{modifier} = \begin{cases} 1.0 \text{ if Gate } > 0.5 \text{ (in-zone)} \\ 0.5 \text{ if Gate } \leq 0.5 \text{ (out-zone)} \end{cases}$$

Function 2: Source-conditional expectations Harris et al. (2013) show that people hold sources to different evidential standards based on perceived expertise. When a knowledgeable source provides surprisingly weak evidence, listeners may draw a pragmatic inference (Grice, 1975) that if this is the best evidence available from someone who should know, the hypothesis must not be well-supported. This inference follows from Grice's maxim of quantity, where cooperative communicators should provide as much information as is required, neither more nor less.

The concept of source-conditional expectations extends Bovens and Hartmann's (2003) framework by making expectations about evidence quality explicit. We do not simply ask "Is this good evidence?" but rather "Is this good evidence *given who is providing it*?" An argument that would be impressive from a novice may be disappointing from an expert, and this disappointment is informative. In the model, expected argument strength scales with source reliability:

$$AS_{expected} = AS_{base} + AS_{rel}\, x\, P(R)$$

This captures two intuitions. First, that we expect some minimum evidence quality from any source attempting to be persuasive (AS_{base}), and second, that we expect higher quality from high-reliability sources (AS_{rel} x P(R)). The deviation from expectation is simply:

$$AS_{dev} = AS - AS_{expected}$$

When AS_{dev} is positive, evidence meets or exceeds expectations, this leads to belief updating toward the speaker position. When $AS_{dev}$is negative, evidence falls short of expectations, triggering the damned by faint praise effect.

Function 3: The Faint Praise Function Following Harris et al (2013), positive arguments can produce negative belief change when they violate epistemic expectations. This follows from pragmatic inferences under uncertainty. When AS_{dev} is non-negative (expectations met), beliefs update toward the speaker position:

$$\text{P(H|rep)} = P(H) + \big(SP - P(H)\big)\; x\; Gate\; x\; w(R)\; x\; |AS_{dev}|^{\beta}$$

However, when AS_{dev} is negative (expectations are not met), beliefs update *away* from the supported position, triggering damned by faint praise:

$$\text{P(H|rep)} = P(H) - \beta_{faint}\; x\; Gate\; x\; w(R)\; x\; |AS_{dev}|^{\beta}\; x\; sign(SP - P(H))$$

β_{faint} is the magnitude of the backfire, which distinguishes GRASP predictions from weighted averaging models. The output P(H|rep) is bounded to [0,1] in implementation, where clipping is invoked only at extreme parameter combinations.

Function 4: Reliability Weighting Bovens and Hartmann (2003) formalised how reliability should rationally affect evidential weight. More reliable sources provide more diagnostic evidence, as their claims carry more information about the true state of the world. The GRASP model extends this framework by distinguishing between noisy and adversarial unreliable sources (Young & de-Wit, 2025).

A noisy source lacks knowledge (uninformative) while an adversarial source has incentives to mislead (triggering reverse inference). The λ-parameter captures this distinction, though the current study tests only noisy sources. For sources above a reliability threshold:

$$w(R) = \left(\frac{P(R) - R_{threshold}}{1 - R_{threshold}}\right)^{\gamma}$$

For sources below threshold, the adversarial λ-parameter determines treatment:

$$w(R) = -\lambda\, x \left(\frac{R_{threshold} - P(R)}{R_{threshold}}\right)^{\gamma}$$

When $\lambda = 0$, unreliable sources are treated as noisy and effectively ignored. When $\lambda > 0$, they are treated as partially or fully adversarial, triggering reverse updating. The γ exponent produces diminishing returns, reflecting that moving from moderate to high reliability has smaller marginal effects than moving from low to moderate.

Function 5: Bi-directional reliability updating Hahn et al. (2009) show that argument quality can provide evidence about source reliability. This may be asymmetric. A weak argument from a trusted source damages credibility (the trust was apparently misplaced). In contrast, weak arguments from already-distrusted sources confirm what was already believed and produce smaller adjustments. This "tall poppy" effect reflects rational inference under uncertainty. Sources incur an implausibility cost when making out-zone claims:

$$Implausibility\ Cost = \alpha_{impl}\, x\, (1 - Gate)\, x\, P(R)$$

Reliability updates based on argument quality, with asymmetric effects when arguments meet expectations:

$$\Delta P(R)_{AS} = \alpha_R \text{ x } AS_{dev} \text{ x } Gate_{modifier} \text{ x } (1 - P(R))$$

While this occurs when arguments are below expectations:

$$\Delta P(R)_{AS} = -\alpha_R\, x\, |AS_{dev}|\, x\, (1.5 - Gate_{modifier}) x\, P(R) x\, \kappa$$

The (1 - P(R)) term provides a ceiling effect, as already trusted sources have less room to gain. The P(R) x κ term implements the "tall poppy" effect where trusted sources have more to lose, with $\kappa > 1$ amplifying this penalty. The reliability given the report is, bounded to [0, 1], computed as:

$$P(Rel|rep) = P(R) + \Delta P(R)_{AS} - Implausibility\ Cost$$

In sum, GRASP's five functions implement one principle: evidence is evaluated relative to source- and argument-conditional expectations. Gating determines whether updating occurs, which is evaluated against prior confidence. Expectations set the evidential bar where faint praise reverses updates when evidence disappoints. Reliability weighting scales magnitude, and bidirectional updating closes the loop between argument quality and source credibility.

Default parameters and core predictions

Default parameters implement moderate, theoretically motivated values: z_threshold = 1.80 SDs (~93% in-zone) approximates the pilot Confidence × Zone manipulations, the reliability exponent γ = 0.70 produces sub-linear weighting consistent with Hahn et al. (2009), and the asymmetric reliability update (κ = 1.50, β_faint = 1.20) ensures faint-praise effects are non-negligible relative to forward updating. These remain a priori values; future work should fit them to data once the qualitative architecture is validated. Table 1 presents the parameters.

Table 1: Default parameters

PARAMETER	DESCRIPTION	DEFAULT
K	Gate steepness	6.0
$z_{threshold}$	Plausibility threshold (in SDs)	1.80
$R_{threshold}$	Minimum reliability for influence	0.30
γ	Diminishing returns exponent	0.70
β	Argument strength exponent	0.50
AS_{base}	Baseline expected AS	0.35
AS_{rel}	Reliability coefficient for expectations	0.25
β_{faint}	Faint praise backfire coefficient	1.20
α_R	Reliability learning rate	0.20
α_{impl}	Implausibility penalty rate	0.15
κ	Reliability penalty amplifier	1.50

We can derive nine predictions from the model for empirical testing. These predictions span belief (P1-4) and reliability updating (P5-9), providing a test of the GRASP architecture.

P1: When AS > $AS_{expected}$, strong arguments move beliefs toward the speaker's position.

P2: Weaker-than-expected arguments trigger faint praise effects where beliefs move away from the speaker's position.

P3: Reliability amplifies argument strength effects. The w(R) term scales all updating by source reliability.

P4: $AS_{expected}$ increases with P(R), so high-reliability sources face higher thresholds and thus larger negative deviations when providing weak evidence.

P5: Stronger-than-expected increase perceived reliability.

P6: Weaker-than-expected decrease perceived reliability.

P7: Argument strength effects dominate initial reliability effects. As arguments provide direct evidence of competence, AS_{dev} should have larger effects than P(R) priors.

P8: High-reliability sources lose more credibility from weak arguments than low-reliability sources.

P9: Zone (plausibility) modulates reliability sensitivity to argument strength. The $Gate_{modifier}$ affects how strongly argument quality updates reliability assessments.

Testing the model: Study 1

Two pilot studies (N = 97, N = 60) calibrated stimuli. Four topics spanned familiar and novel domains to reflect different attitude strength levels: (A) Exercise and Sleep Quality, (B) Handwriting versus Typing for Learning, (C) Interleaving Method for Mathematics, and (D) Meridian Hiring Assessment (fictional). Topics A and B were expected to

elicit confident prior beliefs based on widespread knowledge; Topics C and D were novel claims about which participants would be uncertain implying weaker prior attitude strength.

Strong arguments yielded μ = 72.9% (range 67-84%) and weak arguments rated μ = 10.9% (range 8-16%). High reliability sources were rated μ = 92.4% (range 89-94%) while low sources were rated μ = 11.9% (range 11-13%). Familiar topics elicited higher prior beliefs (μ = 77.5%) than novel topics (μ = 51.8%), t(155) = 8.42, p < .001. Familiar topics also showed narrower credible intervals (μ_σ = 0.085) than novel topics (μ_σ= 0.111), t(155) = 2.31, p < .05, indicating greater confidence.

GRASP predicts a Confidence × Zone interaction where high-confidence participants should resist out-zone claims (gate closed), while low-confidence participants update substantially—a unique signature of gating.

600 participants were recruited via Prolific (age 18+ with English as their native language). After exclusions for failed attention checks (5%) and incomplete responses (4%), 591 participants remained for analysis (μ_{age} = 34.2 years, σ = 11.8; 52% female, 46% male, 2% other/prefer not to say). Participants were randomly assigned to one of eight between-subjects conditions (see below) with approximately 74 participants per cell. Sample size was determined a priori to achieve >88% power for detecting Zone x Reliability interaction (f = 0.15) and >99% power for main effects based on pilot effect sizes.

Materials and procedure

To test the model intuitions, we used a 2 (Confidence: High/Low) x 2 (Zone: In/Out) x 2 (Reliability: High/Low) x 2 (Argument Strength: Strong/Weak) mixed factorial design. Confidence, Zone, and Reliability were between-subjects factors; Argument Strength was within-subjects. This design yields 8 between-subjects cells, each receiving both Strong and Weak argument conditions across two trials.

Topics High Confidence saw Topic A (Exercise) and B (Handwriting) while Low Confidence saw Topics C (Interleaving Method) and D (Meridian Assessment). Low Confidence received brief background information on novel topics before the main trials to ensure basic comprehension.

Zone Speaker Position (SP) was pilot-calibrated. For familiar topics, In-Zone (SP = 75%) aligned with priors while Out-Zone (SP = 20-25%) contradicted them. For novel topics, both In-Zone (SP = 60%) and Out-Zone (SP = 85%) argued FOR the hypothesis, with Out-Zone representing extreme claims.

Reliability Manipulation. High-reliability sources were domain experts with relevant credentials and institutional affiliations. Low-reliability sources were non-experts without relevant credentials, such as anonymous online commenters. They were described as noisy rather than adversarial.

Argument Strength. Strong arguments reported factual sounding percentages, sample sizes, and research citations. Weak arguments were vague, relied on anecdote, and lacked substantive evidential support. As an example, a strong argument related to the hypothesis 'Regular moderate exercise improves sleep quality for most adults' was:

"Based on my systematic review of 47 randomized controlled trials involving over 8,000 participants, I estimate there is approximately a 75% probability that regular moderate exercise improves sleep quality. The pooled data shows that people who exercise regularly fall asleep 12 minutes faster on average and experience 45 minutes more deep sleep per night. The effects are consistent across age groups from 25 to 70. Mechanistic studies show exercise increases adenosine accumulation and regulate core body temperature rhythms. However, I should note that timing matters—exercise within 2 hours of bedtime can actually impair sleep in some people. Effects are also smaller for those who already have good sleep habits"

Each trial presented the hypothesis, source description, argument text, and 0-100 scales for posterior belief and source reliability. Belief change was ΔH = P(H|Rep) – P(H), with positive values indicating movement toward the hypothesis. Reliability was measured asking "How reliable do you think this source is?" Percentage toward SP = (Posterior - Prior) / (SP - Prior), bounded to [0, 100].

Results

Linear mixed-effects models (REML estimation, Satterthwaite degrees of freedom) account for the repeated-measures structure of the design (Argument Strength as a within-subjects factor). Models included random intercepts for participants and fixed effects for all between-subjects factors (Confidence, Zone, Reliability) and their interactions with the within-subjects factor. All factors were contrast-coded (−0.5, +0.5) to facilitate interpretation of main effects and interactions. Effect sizes are reported as Cohen's *d* for pairwise comparisons and unstandardized regression coefficients (*b*) for interaction effects

A linear mixed-effects model predicting belief change revealed significant main effects of Confidence, $b = -15.01$, $SE = 1.48$, $z = -10.15$, $p < .001$, and Argument Strength, $b = 14.66$, $SE = 1.31$, $z = 11.22$, $p < .001$. The main effect of Reliability was also significant, $b = 4.37$, $SE = 1.48$, $z = 2.96$, $p = .003$, indicating that high-reliability sources produced more positive belief change overall. The main effect of Zone was not significant, $b = -2.58$, $SE = 1.48$, $z = -1.74$, $p = .081$.

Several theoretically important interactions emerged. The Confidence × Zone interaction was significant (b = −28.71, z = −9.71, p < .001), consistent with gating. The Zone × Argument Strength interaction (b = −17.55, z = −6.71, p < .001) indicated argument effects were modulated by plausibility. The Confidence × Argument Strength interaction was also significant (b = −8.84, z = −3.38, p = .001). The four-way interaction was significant (b = −20.84, z = −1.99, p = .046), indicating the full pattern depended on all factors.

Belief predictions (P1-4) For strong arguments, 71.7% of participants moved their beliefs closer to the speaker's position, significantly exceeding chance (binomial test, $p < .001$), which supports P1. In conditions where speakers argued for hypotheses with weak evidence, beliefs decreased by $M = -13.77$ points ($SD = 26.43$) on the 100-point scale,

$t(441) = -10.96$, $p < .001$, $d = -0.52$. This negative change from ostensibly supportive evidence supports the pragmatic inference mechanism central to the GRASP model: when evidence falls below expectations, people infer that the hypothesis must not be well-supported, which supports P2.

Reliability × Argument Strength interaction approached significance in the mixed model, $b = 4.79$, $SE = 2.61$, $z = 1.83$, $p = .067$. While the interaction did not reach conventional significance levels, the pattern was in the predicted direction, with high-reliability sources showing amplified sensitivity to argument quality, but P3 was not supported.

Contrary to prediction, the faint praise effect did not differ significantly between high-reliability ($M = -13.91$, $SD = 27.83$) and low-reliability sources ($M = -13.63$, $SD = 24.94$) when tested directly in the subset of FOR conditions, $t(440) = -0.11$, $p = .911$, $d = -0.01$. This null result suggests that in this sample, participants did not hold experts to substantially higher evidential standards than non-experts for belief updating purposes, though both source types did produce significant backfire effects. As such, P4 was not supported.

Reliability predictions (P5-9) A linear mixed-effects model predicting reliability ratings revealed a highly significant main effect of Argument Strength, $b = 35.22$, $SE = 1.21$, $z = 29.01$, $p < .001$, indicating that strong arguments produced substantially higher reliability ratings than weak arguments. The main effect of initial Reliability was also significant, $b = 19.20$, $SE = 1.43$, $z = 13.43$, $p < .001$, showing that sources initially described as high-reliability maintained higher ratings overall.

The Reliability × Argument Strength interaction was significant, $b = 13.46$, $SE = 2.43$, $z = 5.54$, $p < .001$, indicating that argument quality affected reliability ratings differently depending on initial source reliability. The Zone × Argument Strength interaction was also significant, $b = -17.64$, $SE = 2.43$, $z = -7.27$, $p < .001$, showing that the effect of argument strength on reliability judgments was modulated by claim plausibility.

Strong arguments increased reliability (M = 61.27) compared to weak arguments (M = 25.98), t(590) = 26.44, p < .001, d = 1.09 (P5). High-reliability sources dropped from ~91% to 32% after weak arguments, which is a 59-point decline demonstrating that single weak arguments devastate established credibility (P6). Argument strength effects (Δ = 35.29) were 1.82× larger than initial reliability effects (Δ = 19.36), confirming that direct competence evidence outweighs prior reputation (P7; Bovens & Hartmann, 2003).

The significant Reliability × Argument Strength interaction on reliability ratings ($b = 13.46$, $z = 5.54$, $p < .001$) supported the "tall poppy" prediction (P8). High-reliability sources showed a larger drop from strong to weak arguments (Δ = 41.84) than low-reliability sources (Δ = 28.58), a difference of 13.25 points. Trusted sources have more credibility to lose, and participants appropriately penalized them more severely for disappointing argument quality.

The significant Zone × Argument Strength interaction ($b = -17.64$, $z = -7.27$, $p < .001$) indicated that the effect of argument strength on reliability ratings was larger for in-zone claims (Δ = 44.01 points) than for out-zone claims (Δ = 26.54 points), a difference of 17.47 points. This suggests that implausible claims partially protect sources from the full impact of argument quality judgments, perhaps because the claim's implausibility itself provides an alternative explanation for weak arguments. This supports P9. Table 2 summarises the results against model predictions.

Table 2: Predictions versus observations

	PREDICTION	RESULT
P1	Strong AS toward SP	Supported***
P2	Weak AS faint praise backfire	Supported***
P3	Reliability amplifies AS	Marginal (p = .067)
P4	High-reliability larger faint praise	Not supported
P5	Strong AS yields reliability boost	Supported***
P6	Weak AS yields reliability drop	Supported***
P7	AS dominates manipulation	Supported***
P8	High reliability loses more	Supported***
P9	Zone modulates AS on reliability	Supported***

****p < .001.*

The overall pattern provides support for the GRASP model's core mechanisms. The faint praise effect (P2) was robustly confirmed, with weak arguments from sources arguing FOR hypotheses producing significant negative belief change. The reliability updating predictions (P5–P9) were all supported, demonstrating that argument quality bidirectionally affects source credibility assessments. The "tall poppy" effect (P8) and the dominance of argument quality over initial manipulation (P7) confirm that participants used argument quality as direct evidence of source competence.

The two predictions that did not reach conventional significance (P3 and P4) both involved the Reliability × Argument Strength interaction on belief change. While the patterns were directionally consistent with predictions, the effect sizes were smaller than anticipated. This may reflect measurement noise from using aggregated pilot priors rather than individual priors. Alternatively, participants may apply similar evidential thresholds across source types when updating beliefs while calibrating expectations by source expertise when evaluating sources (cf. significant P8).

Discussion and concluding remarks

The GRASP model integrates five mechanisms within a single computational architecture, drawing on Bayesian source-reliability accounts while extending them with non-Bayesian gating and pragmatic-inference components: confidence-based gating, source-conditional expectations, faint praise inference, reliability weighting, and bidirectional reliability updating. Seven of nine predictions received strong empirical support, demonstrating that these mechanisms operate jointly in belief revision. The results point to broader discussions.

Relationship to the Elaboration Likelihood Model

GRASP complements the Elaboration Likelihood Model (Petty & Cacioppo, 1986; Petty, Cacioppo, & Goldman, 1981), which distinguishes central route (argument scrutiny) from peripheral route (source cues) processing. GRASP

formalises this, as gating captures when arguments receive processing and reliability weighting captures source influence. Crucially, GRASP specifies their interaction, as faint praise shows that during central processing, source credibility sets evaluative standards. This extends the ELM by predicting backfire effects from credible sources presenting weak arguments, a pattern unexplained by dual-route frameworks.

The Faint Praise Effect

The robust confirmation of faint praise (P2) supports that Gricean pragmatic inferences operates within belief updating, which is consistent with Harris et al. (2013). A reactance interpretation is difficult to reconcile, as reactance increases with persuasive intent (Brehm, 1966), which predicts stronger backfire from experts. Yet both source types produced equivalent effects. Moreover, faint praise occurred specifically when sources argued FOR hypotheses with weak evidence. Here, a reactance account predicts backfire whenever arguments are weak, regardless of direction. The pattern supports participants inferring what evidence should have been available, which is the pragmatic inference that GRASP formalises.

Source-Conditional Expectations: A Partial Picture

The non-significant Reliability × Argument Strength interactions on belief change (P3, P4) suggest that participants may apply similar evidential thresholds across source types when updating beliefs, even though they differentiate sources when evaluating reliability (significant P8). The faint praise effect (P2) was equally strong for both source types, indicating that pragmatic inference operated but was less source-dependent than predicted.

Bidirectional Reliability Updating

The strong support for reliability updating predictions (P5–P9) contributes to growing evidence that argument quality feeds back onto source evaluations (Hahn et al., 2009; Collins et al., 2018). Argument effects were 1.82 times larger than initial reliability manipulations, consistent with arguments providing diagnostic evidence of competence (Bovens & Hartmann, 2003).

The Zone × Argument Strength interaction (P9) further suggests that implausible claims partially shield sources from credibility damage, as weak arguments for out-zone positions may be attributed to the position's difficulty rather than source incompetence.

Limitations

Several limitations warrant acknowledgement. First, confidence was manipulated via topic familiarity, confounding confidence with domain knowledge. The Confidence × Zone crossover (resistance to out-zone claims only for confident participants) is not predicted by topic differences alone, but future research should manipulate confidence orthogonally.

Second, belief change was computed using aggregated prior beliefs from pilots rather than individual participant priors. This approach avoided demand characteristics from repeated measurement but introduced measurement uncertainty that likely attenuates interaction effects, possibly contributing to the P3 and P4 null results.

Third, we tested qualitative predictions rather than fitting the model quantitatively, which follows recommendations for theory-driven testing that avoids overfitting (Guest & Martin, 2021; Wilson & Collins, 2019). Supplementary calibration analyses (available from authors) confirm that optimised parameter settings preserve all effects and that the model with these parameters enjoys a strong quantitative fit with posteriors beliefs ($R^2 > .85$).

Fourth, the distinction between noisy and adversarial sources (Function 4) was not directly tested. The low-reliability sources were noisy rather than adversarial. This mechanism awaits testing in contexts with motivated sources.

Fifth, our nine directional predictions were derived a priori from the model's functional structure but were not formally pre-registered; replication with pre-registration would strengthen confidence in the architecture. Finally, future research should measure individual priors, manipulate confidence independently, employ continuous factor levels, and explore individual differences. The model could be extended to address misinformation contexts (Connor Desai et al., 2020; Ecker et al., 2022).

Conclusion

GRASP demonstrates that multiple belief revision phenomena emerge from a common framework: evidence is evaluated relative to source- and argument-conditional expectations. The partial support for source-conditional expectations in belief updating, alongside strong support for this mechanism in reliability updating, suggests that people may apply these expectations asymmetrically: more for evaluating *who to trust* than for deciding *what to believe*.

References

Birnbaum, M. H., & Mellers, B. A. (1983). Bayesian inference: Combining base rates with opinions of sources who vary in credibility. *Journal of Personality and Social Psychology*, 45(4), 792–804.

Bovens, L., & Hartmann, S. (2003). *Bayesian epistemology*. Oxford University Press.

Brehm, J. W. (1966). *A theory of psychological reactance*. Academic Press.

Collins, P. J., Hahn, U., von Gerber, Y., & Olsson, E. J. (2018). The bi-directional relationship between source characteristics and message content. *Frontiers in Psychology*, 9, 18.

Connor Desai, S. A., Pilditch, T. D., & Madsen, J. K. (2020). The rational continued influence of misinformation. *Cognition*, 205, 104453.

Coppock, A. (2022). *Persuasion in parallel: How information changes minds about politics*. University of Chicago Press.

Corner, A., Hahn, U., & Oaksford, M. (2011). The psychological mechanism of the slippery slope argument. *Journal of Memory and Language*, 64(2), 133–152.

Ecker, U. K. H., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., Kendeou, P., Vraga, E. K., & Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, 1(1), 13–29.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics:* Vol. 3. Speech acts (pp. 41–58). Academic Press.

Guest, O., & Martin, A. E. (2021). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science*, 16(4), 789–802.

Hahn, U., Harris, A. J. L., & Corner, A. (2009). Argument content and argument source: An exploration. *Informal Logic*, 29(4), 337–367.

Hahn, U., & Oaksford, M. (2007). The rationality of informal argumentation: A Bayesian approach to reasoning fallacies. *Psychological Review*, 114(3), 704–732.

Hahn, U., Oaksford, M., & Harris, A. J. L. (2012). Testimony and argument: A Bayesian perspective. In F. Zenker (Ed.), *Bayesian argumentation* (pp. 15–38). Springer.

Harris, A. J. L., Corner, A., & Hahn, U. (2013). James is polite and punctual (and useless): A Bayesian formalisation of faint praise. *Thinking & Reasoning*, 19(3–4), 414–429.

Harris, A. J. L., Hahn, U., Madsen, J. K., & Hsu, A. S. (2016). The appeal to expert opinion: Quantitative support for a Bayesian network approach. *Cognitive Science*, 40(6), 1496–1533.

Klayman, J., & Ha, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, 94(2), 211–228.

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, 13(3), 106–131.

Madsen, J. K. (2019). Voter reasoning bias when evaluating statements from female and male election candidates. *Politics & Gender*, 15(2), 310–335.

Madsen, J. K., Hahn, U., & Pilditch, T. D. (2016). The impact of partial source dependence on belief and reliability revision. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 2330–2335). Cognitive Science Society.

Madsen, J. K., Hahn, U., & Pilditch, T. D. (2020). The impact of partial source dependence on belief and reliability revision. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 46(9), 1795–1805.

Mercier, H. (2020). *Not born yesterday: The science of who we trust and what we believe*. Princeton University Press.

Merdes, C., von Sydow, M., & Hahn, U. (2021). Formal models of source reliability. *Synthese*, 198(Suppl 23), 5773–5801.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press.

Olsson, E. J. (2013). A Bayesian simulation model of group deliberation and polarization. In F. Zenker (Ed.), *Bayesian argumentation* (pp. 113–133). Springer.

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. Advances in Experimental *Social Psychology*, 19, 123–205.

Petty, R. E., Cacioppo, J. T., & Goldman, R. (1981). Personal involvement as a determinant of argument-based persuasion. *Journal of Personality and Social Psychology*, 41(5), 847–855.

Petty, R. E., & Krosnick, J. A. (Eds.). (1995). *Attitude strength: Antecedents and consequences.* Lawrence Erlbaum Associates.

Pilditch, T. D., Hahn, U., & Lagnado, D. (2025). The problem of dependency. *Synthese*, *205*(4), 143.

Rollwage, M., Loosen, A., Hauser, T. U., Moran, R., Dolan, R. J., & Fleming, S. M. (2020). Confidence drives a neural confirmation bias. *Nature Communications*, 11, 2634.

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language*, 25(4), 359–393.

Tormala, Z. L., & Petty, R. E. (2004). Resistance to persuasion and attitude certainty: The moderating role of elaboration. *Personality and Social Psychology Bulletin*, 30(11), 1446–1457.

Visser, P. S., & Mirabile, R. R. (2004). Attitudes in the social context: The impact of social network composition on individual-level attitude strength. *Journal of Personality and Social Psychology*, 87(6), 779–795.

Wilson, R. C., & Collins, A. G. E. (2019). Ten simple rules for the computational modeling of behavioral data. *eLife*, 8, e49547.

Young, D. J., & de-Wit, L. H. (2025). The bias-and-expertise model: A Bayesian network model of political source characteristics. *Cognitive Science*, 49(1), e70009.

Young, D. J., Madsen, J. K., & De-Wit, L. H. (2025). Belief polarization can be caused by disagreements over source independence: Computational modelling, experimental evidence, and applicability to real-world politics. *Cognition*, *259*, 106126.