Coherence and Bayesian Reasoning in Legal Judgment: How Verdicts Reshape Evidence Evaluation

Mengxuan Helen Qiao*[1] (helen.qiao.18@ucl.ac.uk)

David G. Kamper*[2] (davidgkamper@ucla.edu)

David A. Lagnado[1] (d.lagnado@ucl.ac.uk)

Keith J. Holyoak[2] (kholyoak@g.ucla.edu)

[1]Department of Experimental Psychology, University College London

[2]University of California, Los Angeles

* These authors contributed equally.

Abstract

Coherence-based theories propose that decision-makers actively reinterpret evidence to align with their emerging conclusions. Prior research on legal decision-making demonstrated coherence shifts in ratings of agreement with rival arguments, but it remains unclear whether similar effects can be found using measures more directly related to normative evidential strength. Across three studies ($N = 339$), participants evaluated prosecution and defense arguments in a mock burglary case before and after rendering verdicts. In Study 1, participants exhibited clear coherence shifts: guilty verdicts were associated with increased agreement with prosecution arguments and decreased agreement with defense arguments, whereas not-guilty verdicts produced the opposite pattern. In Studies 2a–b, a likelihood-based measure of arguments strength also changed across testing periods; however, these shifts were not reliably aligned with verdicts. A constraint satisfaction model was able to predict mean shifts from pre- to post-verdict evaluations, and Bayesian network modeling using participants' post-verdict probability estimates closely approximated empirical guilt judgments. It appears that constraint satisfaction can drive coherence shifts during the process of reaching a verdict, such that verdicts and post-verdict evaluations of evidence are consistent with Bayesian inference.

Keywords: legal decision-making; coherence-based reasoning; constraint satisfaction; Bayesian inference; belief revision

Introduction

How do people reason about evidence when making high-stakes decisions? A normative account assumes that decision-makers objectively weigh evidence, integrate information in accord with formal principles of rationality, and reach conclusions reflecting the balance of evidence. This idealized model underpins many institutional practices, particularly in law, where jury instructions emphasize impartial evaluation, and appellate doctrines presume that jurors reliably discriminate between stronger and weaker evidence.

However, decades of research in psychology suggests a more complex picture. A substantial body of work demonstrates that legal decisions are easily influenced by extra-legal factors that bear no causal relation to guilt (e.g., Cheung & Lagnado, 2023), features of evidentiary presentation and procedural design (e.g., Qiao & Lagnado, 2025), as well as confirmation biases (Nickerson, 1998) and motivated reasoning (Kunda, 1990). These findings raise doubts about whether a Bayesian framework, which has long been treated as a normative benchmark for rational inference, is sufficient to accurately capture how fact-finders evaluate evidence in practice.

Bayesian models typically assume that each evidence item is assessed independently, conditional on the hypotheses. Human cognition, however, may operate less like a statistical calculator and more like a connectionist system: a *constraint satisfaction network* (CSN), in which different pieces of information are integrated to construct a coherent explanatory account, even at the cost of distorting the perceived diagnostic value of evidence to better fit that account.

This process, termed *coherence-driven reasoning* (Holyoak & Simon, 1999; Simon et al., 2015; Spellman et al., 1993; Thagard, 1989), involves bidirectional influence between conclusions and evidence: not only does evidence shape conclusions, but conclusions simultaneously reshape evidence interpretation. This process results in a coherent belief system that may deviate from normative Bayesian standards. A critical implication of this process is that the assumed independence between evidence items, the evaluations of evidence, and verdicts breaks down. For example, a juror who initially views fingerprint evidence as moderately probative might, after developing a preliminary inclination toward guilt that has been informed by other pieces of evidence, come to view that same evidence as more compelling. Such feedback loops potentially entrench erroneous judgments and, in turn, threaten core due process values, including impartial evaluation and the presumption of innocence.

Given the conceptual tension and empirical ambiguity, this paper examines two influential yet conceptually distinct frameworks for understanding legal decision-making: Bayesian reasoning, with particular focus on Bayesian networks (BNs), and coherence-based reasoning. We first outline the theoretical foundations of each approach and their respective accounts of evidence evaluation and verdict formation. We then present empirical studies testing whether coherence shifts occur when evidence is evaluated using Bayes' factors, as it does when evaluated by agreement ratings. We further present BN and CSN models fitted to our empirical

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

data, assessing the predictive accuracy of each model and the extent to which Bayesian probabilistic inference may be captured within—or be compatible with—a coherence-driven process. Finally, we conclude by discussing the implications of our findings for accounts of juror reasoning.

Bayesian Network

Bayes' theorem provides mathematical rules for updating beliefs about a hypothesis in light of observed evidence, specifying how the probability of a cause may be inferred from its effect. In the legal context, this process commonly begins with a prior belief about guilt denoted as $P(G)$. Upon observing evidence E, this prior belief is updated to yield a posterior probability of guilt, $P(G|E)$. This updating depends on the likelihood ratio (LR) of E, which captures its probative value by comparing the probability of observing E if the guilty hypothesis is true, $P(E|G)$, with the probability of observing E if the hypothesis is false, $P(E|\neg G)$ (Lagnado, 2021). Although this inferential process may not be explicitly articulated by lay decision-makers, empirical research suggests that individuals often approximate Bayesian updating at a qualitative level when revising beliefs about guilt in response to evidence (e.g., Qiao & Lagnado, 2025; Shengelia & Lagnado, 2021).

BNs extend Bayes' theorem to complex inferential settings involving multiple hypotheses, pieces of evidence, and interdependencies. A BN is a graphical model grounded in Bayesian probability theory that is used to visualize and model the causal and probabilistic relationships between hypotheses and pieces of evidence (Fenton et al., 2013; Pearl, 2000). BNs, and closely related probabilistic graphical approaches, have a long history of application in legal reasoning and evidence evaluation (e.g, Fenton & Neil, 2011; Friedman, 1986; Taroni et al., 2006). Structurally, a BN consists of nodes, which represent uncertain variables, and directed links, which encode causal or dependency relationships among those variables. Each node is associated with a conditional probability table (CPT) that specifies the probability of the effect variable given each configuration of its direct parent nodes. Together, the network and CPTs allow for a rational, structured transition from prior beliefs to an updated verdict by explicitly modeling the ramification of different assumptions and pieces of evidence.

Coherence-Based Reasoning

An alternative framework describes the legal reasoning process as constructing coherent narratives of the alleged events (Pennington & Hastie, 1988). In this *coherence-based reasoning* account, individual pieces of evidence are interpreted and weighted according to their contribution to the overall explanatory coherence. Holyoak and Simon (1999) argue that coherence-based reasoning involves bidirectional influence between conclusions and evidence, operating as a parallel constraint satisfaction process within a CSN. Evidence items and potential verdicts correspond to nodes in a network, with connection weights reflecting mutual consistency or inconsistency. The network iteratively adjusts activations until reaching a stable state in which mutually consistent elements are strongly activated and inconsistent elements suppressed. Critically, this process transforms evidence interpretation: ambiguous facts are interpreted in the way that best fits the emerging narrative.

Empirical support for coherence shift comes from studies with mock jurors, which have shown that evidence evaluations shift toward consistency with verdict preferences (Holyoak & Simon, 1999; Simon, 2004). These studies used pre–post designs in which participants rated their agreement with prosecution and defense arguments before and after reaching a verdict. Such agreement ratings reflect subjective endorsement rather than normative evidential weight, leaving an open question as to whether the observed shifts represent a departure from Bayesian inference.

Coherence shifts bear some resemblance to biased assimilation, whereby individuals interpret ambiguous evidence in ways that align with pre-existing beliefs, potentially producing belief polarization—i.e., people with opposing priors observe the same evidence yet each strengthen their original position (Lord et al., 1979). Coherence shifts differ in that they do not presuppose differences in prior beliefs; rather, it is the emerging conclusion, formed during evidence evaluation, that feeds back to shape evidence evaluation. Nevertheless, the two phenomena share the implication of apparent irrationality. Crucially, however, Jern et al. (2014) demonstrate that belief polarization can be predicted by normative probabilistic inferences with a BN, suggesting that what appears irrational can be compatible with normative accounts.

The present studies build on this insight in two respects. First, we examine whether coherence shifts extend beyond subjective agreement ratings to LRs—a formally-grounded measure of evidential weight that is the central diagnostic quantity in Bayesian inference. Second, we test whether the operation of a CSN can result in a verdict and set of beliefs about evidence consistent with Bayesian probabilistic inference, implying that coherence-driven reasoning and Bayesian reasoning may converge.

Overview of Studies

We investigate coherence-driven belief revision using a mock burglary case with balanced prosecution and defense arguments. Participants evaluate these arguments both before and after seeing the case and rendering a verdict.

Study 1, a conceptual replication of Holyoak and Simon (1999), employed agreement ratings assessing endorsement of prosecution and defense claims. Studies 2a–b assessed conditional probabilities (CPs): $P(A|G)$ and $P(A|\neg G)$. From these elicited values, we computed log LRs that quantify evidence strength. Study 1 and Study 2a used a fixed order of argument presentation, whereas Study 2b randomized presentation order to control for potential order effects.

A central prediction for coherence-based reasoning is that from pre- to post-test, guilty verdicts should increase agreement with prosecution arguments and decrease agreement with defense arguments, with an opposite pattern for not-guilty verdicts. We applied a CSN to predict mean shifts in evaluations from pre- to post-verdict, and constructed a BN to assess the normative consistency between verdicts and evidence evaluations.

All materials, data, and analyses are available at the OSF repository (https://osf.io/r5cbf).

Study 1: Agreement Ratings

Method

Participants We recruited 130 participants through Prolific Academic from a U.S. sample ($N_{\text{female}} = 69$, $N_{\text{male}} = 60$, $N_{\text{other}} = 1$; $M_{\text{age}} = 43.4$, $SD_{\text{age}} = 12.9$). Participants were pre-screened for English fluency and 95–100% platform approval rate. All provided informed consent and were reimbursed $1.96.

Design The experiment employed a 2 (Verdict: guilty vs. not-guilty; between-subjects) × 2 (Test: pre vs. post; within-subjects) × 2 (Side: prosecution vs. defense; within-subjects) × 4 (Evidence; within-subjects) mixed design, where Verdict was determined by participants' own judgments.

Materials and Procedure

Pre-test Measures Participants first rated their agreement with eight general statements about burglary, conceptually aligned with the prosecution and defense arguments they would see later (e.g., *'How much do you agree that, in a burglary case, a fingerprint matching a suspect's record would be found at the point of entry, establishing that they touched the exact location where the burglar entered?'*). Ratings were made on a 0 (not at all) to 100 (completely) scale. The argument presentation order was held constant throughout the study, with prosecution arguments for evidence items 1–4 presented before the corresponding defense arguments. These ratings established baseline endorsement levels prior to case exposure and verdict formation.

Case Presentation and Verdict Participants then read a mock burglary case in which Derek, a 22-year-old repair technician, was accused of burglarizing Vivian's home. Four pieces of evidence were presented—fingerprint match, fiber match, security camera footage, and internet search history, each of which could support either verdict depending on interpretation. For example, the fingerprint could indicate Derek touched the entry point during the burglary, or that it was left during a legitimate prior repair visit. As Lagnado (2021) notes, such evidence carries probative weight only when embedded in a coherent causal account; its value is diminished when compatible with an innocent explanation. Each piece of evidence was accompanied by one prosecution and one defense argument, yielding eight arguments in total.

Following the case, participants provided (1) a probability of guilt judgment from 0% (impossible) to 100% (certain), and (2) a binary verdict decision (guilty or not-guilty).

Post-test Measures Participants re-rated their agreement with each of the eight arguments as instantiated in the case (e.g., *'How much do you agree that a fingerprint matching Derek's was found at the entry point, establishing that he touched the exact location where he entered during the burglary?'*) on the 0 (not at all) to 100 (completely) scale.

Results and Discussion

Of 130 participants, 74 rendered guilty verdicts and 56 not-guilty verdicts. Mean estimate of guilt probability was .80 among guilty deciders and .41 among not-guilty deciders.

We collapsed individual arguments within sides and conducted a 2 (Verdict) × 2 (Side) × 2 (Test) mixed ANOVA. It revealed a significant main effect of Test, $F(1,128) = 29.37$, $p < .001$, and Side, $F(1,128) = 47.66$, $p < .001$.

Critically, we found a significant three-way Verdict × Test × Side interaction, $F(1,128) = 96.31$, $p < .001$. On the pre-test, agreement did not differ by verdict ($p > .119$ for both sides). On the post-test, participants who reached a guilty verdict agreed more with prosecution arguments ($M = .80$) than did not-guilty deciders ($M = .53$), $t(128) = 8.36$, $p < .001$; whereas not-guilty deciders agreed more with defense arguments ($M = .72$) than did guilty deciders ($M = .42$), $t(128) = 7.75$, $p < .001$. For guilty deciders, prosecution agreement increased significantly from pre- ($M = .62$) to post-test, $t(128) = 5.75$, $p < .001$. For not-guilty deciders, defense agreement increased from pre- ($M = .44$) to post-test, $t(128) = 9.22$, $p < .001$ (see Figure 1).

These results replicate those of Holyoak and Simon (1999), demonstrating that participants' evaluations of arguments shifted to cohere with their emerging verdicts: before deciding, agreement did not predict verdict; after deciding, evaluations diverged systematically.

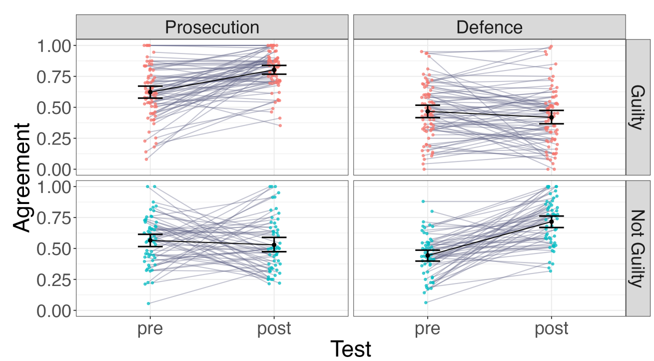

Figure 1: Mean agreement scores in Study 1 for prosecution and defense arguments by verdict and test phase. Each line connects ratings from an individual participant; heavy line indicates mean values. Error bars show ±1 SE.

Studies 2a–b: Likelihood Ratios

Study 1 found coherence shifts in subjective agreement ratings. In Study 2, we extended this analysis by examining whether coherence shifts emerged when argument strengths were evaluated using LRs. Notably, the conditional probabilities elicited in Study 2 do not correspond directly to Bayesian quantities, which would be the probability of evidence E given either hypothesis (i.e., $P(E|G)$ and $P(E|\neg G)$). Here, participants were asked to estimate the conditional probabilities of arguments A supporting or opposing the evidence E: $P(A|G)$ and $P(A|\neg G)$. For example, $P(E|G)$ would correspond to the probability of a fingerprint match given the guilty hypothesis; whereas $P(A|G)$ corresponds to the probability of the fingerprint match indicating the suspect touching the location where the burglar entered the crime scene given the guilty hypothesis.

Method

Participants Study 2a recruited 126 participants ($N_{\text{female}} = 68$, $N_{\text{male}} = 57$, $N_{\text{prefer not to say}} = 1$; $M_{\text{age}} = 43.4$, $SD_{\text{age}} = 11.8$); Study 2b recruited 83 participants ($N_{\text{female}} = 43$, $N_{\text{male}} = 35$, $N_{\text{other}} = 4$; $N_{\text{prefer not to say}} = 1$; $M_{\text{age}} = 43.6$, $SD_{\text{age}} = 12.8$), with one excluded for completing too quickly (< 2 minutes), leaving 82. No participant completed multiple studies. All provided informed consent and were reimbursed \$1.96.

Design Both studies used the same $2 \times 2 \times 2 \times 4$ design as Study 1.

Materials and Procedure Both studies followed the same procedures as in Study 1. However, instead of agreement ratings, participants evaluated the likelihood of each argument under both guilt and not-guilty hypotheses: $P(A|G)$ and $P(A|\neg G)$. On the pre-test, questions were generic (e.g., '*If someone is guilty / not guilty of burglary, how likely do you think that a fingerprint matching their record would be found at the point of entry, establishing that they touched the exact location where the burglar entered the scene during the burglary?*'); on the post-test, questions referred specifically to the case (e.g., '*If Derek was guilty / not guilty of burglary, how likely do you think that a fingerprint matching their record would be found at the point of entry, establishing that they touched the exact location where the burglar entered the scene during the burglary?*'). Study 2a presented arguments in a fixed order as in Study 1; Study 2b used a randomized order.

Results and Discussion

Study 2a Of 126 participants, 67 voted guilty ($M_{\text{guilt}} = .77$) and 59 not-guilty ($M_{\text{guilt}} = .41$).

To compute LRs, responses of 0 and 1 were recoded to .001 and .999 to prevent division by zero. LRs were subsequently log-transformed, making the scale interpretable and suitable for parametric analysis. As in Study 1, we collapsed individual arguments within sides. A mixed ANOVA on log LRs revealed significant main effects of Verdict, $F(1,124) = 8.96$, $p = .003$; Side, $F(1,124) = 153.98$, $p < .001$; and Test, $F(1,124) = 67.77$, $p < .001$.

There was a significant Verdict $\times$ Test interaction, $F(1,124) = 5.10$, $p = .026$. Pre-test argument strength did not differ by verdict ($p = .58$); on the post-test, guilty deciders judged arguments more indicative of guilt ($M = 1.08$) than did not-guilty deciders ($M = 0.25$), $t(124) = 7.19$, $p < .001$. The three-way interaction approached significance ($p = .083$; see Figure 2).

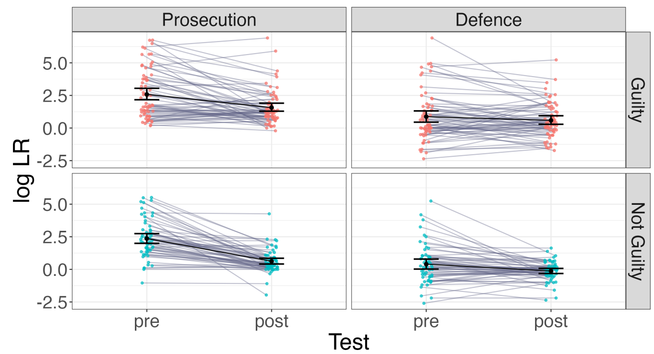

Figure 2: Mean log LRs in Study 2a for prosecution and defense by verdict and test phase. Each line connects ratings from an individual participant; heavy line indicates mean values. Error bars show ± 1 SE.

Study 2b Of 82 participants, 43 voted Guilty ($M_{\text{guilt}} = .79$) and 40 Not Guilty ($M_{\text{guilt}} = .40$).

Results generally replicated Study 2a: we found significant main effects of Side, $F(1,81) = 154.85$, $p < .001$, and Test, $F(1,81) = 34.14$, $p < .001$; and a significant Verdict $\times$ Test interaction, $F(1,81) = 4.98$, $p = .028$. The three-way interaction was not significant ($p = .44$). This replication confirms that order effects did not account for the results of Study 2a. (A figure and pairwise comparisons for Study 2b can be found in the OSF repository.)

Across both Studies 2a and 2b, there was a general decrease of evaluations of both prosecution and defense arguments after reading the case and deciding a verdict, for both guilty and not-guilty deciders. However, the absence of reliable three-way interactions suggests that these changes were not systematically aligned with the specific verdict participants reached. Taken together, these results indicate that when argument strength was operationalized using log LRs, evaluative shifts did not exhibit the same pattern of coherence shift as when they were measured by subjective agreements.

Bayesian Network Modeling

We combined data from Studies 2a–b ($N = 204$: 107 guilty, 97 not-guilty) to parameterize a BN using post-test probability estimates. We could not use pre-test responses to predict posteriors because the phrasing of the questions could not validly reflect the conditional probabilities assumed by BNs.

Model Structure

The network included four evidence nodes (E), each generated either by guilt (G) or an alternative explanation (AE). The structure is illustrated in Figure 3.

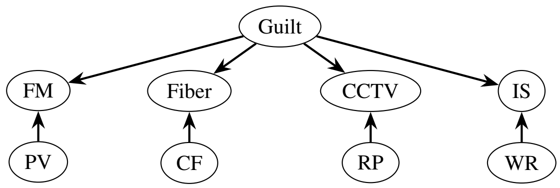

Figure 3: Bayesian network structure. FM = Fingerprint Match; PV = Prior Visit; Fiber = Fiber Evidence; CF = Common Fibers; CCTV = CCTV footage; RP = Routine Path; IS = Internet Search; WR = Work Related.

The estimations of CPTs were partly from participants' responses to the CP questions. The mapping was determined by the wording of the arguments. We translated $P(A|G)$ and $P(A|\neg G)$ for prosecution arguments directly into $P(E|G)$ and $P(E|\neg G)$; while $P(A|G)$ and $P(A|\neg G)$ for defense arguments were translated into $P(E|G,AE)$ and $P(E|\neg G,AE)$. Using the law of total probability, we derived $P(E|G,\neg AE)$ and $P(E|\neg G,\neg AE)$ to complete the CPTs for each evidence node. Priors for $P(AE)$ came from a separate pilot study ($N = 30$; dataset available on OSF). The prior of guilt was set to 0.5. All estimations were taken or calculated from group means of guilty or not-guilty deciders. We conducted the BN modeling using agena.ai (https://www.agena.ai).

Results

Setting all evidence and alternative explanation nodes to true, the model predicted a posterior guilt probability of .707 for guilty deciders (vs. empirical estimate of .779) and .402 for not-guilty deciders (vs. empirical estimate of .405).

Discussion

Based on post-verdict CPs, the BN yielded posterior probabilities of guilt that closely approximated participants' mean reported probabilities. This result suggests that the Bayesian framework captures participants' evidential reasoning relatively well, with differential evaluations of CPs depending on the guilty or not-guilty verdict.

Nonetheless, limitations of our modeling preclude strong conclusions regarding the psychological realism of the Bayesian framework as an account of juror decision-making. First, we assumed that all participants implicitly constructed and relied on an identical BN to explain the evidence by guilt versus alternative explanations. In practice, participants may have differed in how they conceptualized causal relationships among these variables. Second, the estimates for $P(E|G,\neg AE)$ were frequently extremely close to 1. Such estimates are theoretically implausible, as the probability of observing evidence given guilt when the alternative explanation is false should not systematically approach certainty. Moreover, this value should generally be lower than $P(E|G,AE)$, as the additional explanation should increase the likelihood of evidence, which was not the case. This distortion likely arises from the modeling constraint imposed by using the maximum permissible value of $P(AE)$ when computing the CPTs.

Taken together, the BN reproduces probability judgments very closely, although the limitations indicate that the parameter estimates themselves may reflect artifacts of the modeling assumptions. As such, the present results should be interpreted cautiously and further analyses are required using more rigorous parameter estimates at the level of individual participants.

Constraint Satisfaction Model

To provide an account of mean coherence shifts, we implemented a CSN model similar to that used by Holyoak and Simon (1999). The model formalizes the hypothesis that evidence evaluations and verdict decisions mutually constrain each other, settling into coherent configurations through iterative relaxation.

Model Architecture

The network comprises 30 nodes organized hierarchically (see Figure 4): 16 CP nodes representing $P(A|G)$ and $P(A|\neg G)$ for each of eight arguments, 8 agreement nodes, 2 verdict nodes, and 4 control/anchor nodes (S, P, D, X). The special node S provides data-driven input to evidence nodes weighted by pre-test probabilities. Control nodes P and D coordinate prosecution and defense arguments, respectively, with mutual inhibition. The bias node X connects to the chosen verdict, anchoring the network toward observed decisions. The X node can be interpreted as summarizing the unknown factors that led individual participants to reach their observed verdict. The CSN only aims to predict aggregate shifts at the level of groups means for participants who reached each verdict. The CSN was applied to the mean data for participants who decided guilty versus those who decided not-guilty.

Critically, connections from evidence nodes to agreement nodes were *unidirectional*—CP influences agreement, but not vice versa. This architectural choice (essential for model fit) was motivated by the observation that evaluations of CPs showed smaller shifts than agreement ratings. The theoretical interpretation is that evidence evaluations are based on conditionals ('If the defendant is guilty, then...'), but not vice versa. Conditionals reflect a participant's prior beliefs under the hypothesis of guilt or else innocence, whereas agreement ratings are dependent on the conditionals. All other connections are bidirectional, with inhibitory connections were between opposing verdict nodes between prosecution and defense nodes.

Model Fitting

We estimated 11 free parameters via grid search over 100,000 parameter combinations, minimizing summed squared error between predicted and observed post-test values across $P(A|G)$, $P(A|\neg G)$, and agreement ratings. Because Study 1 collected only agreement ratings while Studies 2a–b collected only CPs, we combined group-level data from both. For each verdict group, we computed mean post-test values for each of

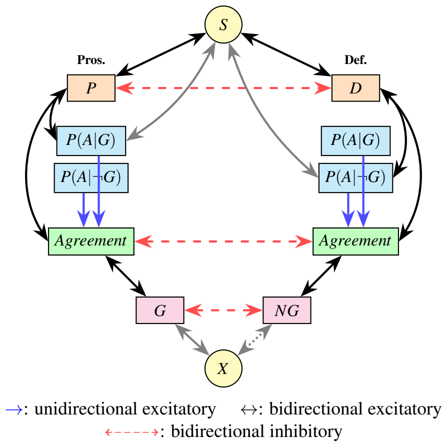

Figure 4: CSN architecture. S = special anchor; P/D = control nodes; X = bias node connecting to chosen verdict.

the 8 arguments, yielding 16 data points per measure (8 arguments $\times$ 2 verdicts). The model thus predicted 48 values in total: 16 for $P(A|G)$, 16 for $P(A|\neg G)$, and 16 for agreement.

Results

The best-fitting model achieved strong correspondence between predictions and observations: $r = .87$ for $P(A|G)$, $r = .79$ for $P(A|\neg G)$, and $r = .86$ for agreements (all $p < .001$). Root mean squared errors ranged from .080 to .093.

A comparison model with bidirectional CP-to-agreement connections performed substantially worse, confirming that coherence operates asymmetrically here: verdicts reshape agreement with argument interpretations, but agreement does not strongly feed back to alter likelihood estimates.

Discussion

The CSN model provides a mechanistic account of coherence-driven belief revision. By representing CPs, agreements, and verdicts as interconnected nodes that settle into mutually consistent states, the model captures the systematic restructuring observed from pre- to post-verdict. The necessity of unidirectional CP-to-agreement connections suggests that coherence primarily operates at the level of subjective agreement rather than probability estimation, consistent with the larger effect sizes observed for agreement ratings in our behavioral analyses.

General Discussion

In Study 1, participants' agreements with arguments systematically shifted in the direction of their verdict from pre- to post-test to support their decision of whether the defendant is guilty or not guilty, replicating the findings from Holyoak and Simon (1999). In Study 2, likelihood-based measures also exhibited changes across testing periods: there was a general decrease in the evaluated diagnosticity of both prosecution and defense arguments from pre- to post-test. However, unlike in Study 1, these shifts did not differ as a function of participants' verdict choices.

The differential magnitude across measures is theoretically informative. Agreement ratings tap narrative endorsement and subjective plausibility, which may be more malleable than formal probability estimates. Likelihood judgments require explicit consideration of alternative hypotheses, potentially promoting analytical reasoning that resists coherence pressures. The CSN model's inclusion of unidirectional CP-to-agreement connections reinforces this interpretation: coherence primarily reshapes evaluative endorsement rather than feeding back to alter CP estimates. It should be noted, however, that the CSN model does predict small coherence shifts at the level of CPs (as opposed to LRs), as was observed in our data from Study 2.

At the level of group means, both the BN and CSN models predicted participants' post-verdict judgments relatively well. Using participants' CP judgments on the post-test, the guilt probability estimates provided by the BN closely approximated empirical judgments of guilt. This finding suggests that legal evidence evaluation may, at least in part, conform to Bayesian principles. Moreover, by separating probability estimation by verdict choices, this analysis indicates that participants evaluated the likelihood of arguments in a systematic manner to inform and support their judgments of guilt. Complementing this normative account of post-verdict belief, the CSN model provided a process-level explanation of how a decision may be reached: CPs, agreements, and verdicts settle into mutually consistent configurations through constraint satisfaction, with the model achieving high correlations with human evaluations on the post-test.

The convergence of predictions across models suggests that, although the two frameworks approach legal reasoning from different theoretical perspectives, they are not necessarily competing or incompatible. Participants adjusted their evaluations of evidence in ways that promoted coherence with emerging verdict, while simultaneously exhibiting inference patterns that were broadly consistent with Bayesian principles. This finding is consistent with Jern et al. (2014)'s demonstration that belief polarization can be captured by normative probabilistic inference. Coherence-driven and Bayesian-consistent processes may operate in tandem during legal decision-making, rather than offering mutually exclusive accounts.

Although we established a robust association between evidence evaluation and verdict formation, the causal direction of this relationship remains unresolved, leaving open questions for future research. Subsequent work should also examine whether individual differences in analytical thinking moderate coherence effects, and whether group deliberation among jurors amplifies or attenuates these biases.

Acknowledgment

We would like to thank Dan Simon for helpful discussions during the early stages of this project.

References

Cheung, V., & Lagnado, D. (2023). Defendant character influences mock juror judgments of blame, guilt, and punishment. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*.

Fenton, N., & Neil, M. (2011). The use of Bayes and causal modelling in decision making, uncertainty and risk. *CEPIS Upgrade*, *12*(5), 10–21.

Fenton, N., Neil, M., & Lagnado, D. A. (2013). A general structure for legal arguments about evidence using Bayesian networks. *Cognitive Science*, *37*(1), 61–102.

Friedman, R. D. (1986). Route analysis of credibility and hearsay. *Yale Law Journal*, *96*, 667.

Holyoak, K. J., & Simon, D. (1999). Bidirectional reasoning in decision making by constraint satisfaction. *Journal of Experimental Psychology: General*, *128*(1), 3–31. https://doi.org/10.1037/0096-3445.128.1.3

Jern, A., Chang, K.-M. K., & Kemp, C. (2014). Belief polarization is not always irrational. *Psychological Review*, *121*(2), 206–224.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480–498. https://doi.org/10.1037/0033-2909.108.3.480

Lagnado, D. A. (2021). *Explaining the evidence: How the mind investigates the world*. Cambridge University Press.

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, *37*(11), 2098–2109.

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

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge University Press.

Pennington, N., & Hastie, R. (1988). Explanation-based decision making: Effects of memory structure on judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(3), 521–533. https://doi.org/10.1037/0278-7393.14.3.521

Qiao, M. H., & Lagnado, D. (2025). Speak last and step-by-step: The effect of order and response mode on evidence evaluation. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *47*.

Shengelia, T., & Lagnado, D. (2021). Are jurors intuitive statisticians? Bayesian causal reasoning in legal contexts. *Frontiers in Psychology*, *11*, 519262.

Simon, D. (2004). A third view of the black box: Cognitive coherence in legal decision making. *University of Chicago Law Review*, *71*(2), 511–586.

Simon, D., Stenstrom, D. M., & Read, S. J. (2015). The coherence effect: Blending cold and hot cognitions. *Journal of Personality and Social Psychology*, *109*(3), 369–394.

Spellman, B. A., Ullman, J. B., & Holyoak, K. J. (1993). A coherence model of cognitive consistency: Dynamics of attitude change during the Persian Gulf War. *Journal of Social Issues*, *49*(4), 147–165.

Taroni, F., Aitken, C. G., Garbolino, P., & Biedermann, A. (2006). *Bayesian networks and probabilistic inference in forensic science*. Wiley Online Library.

Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences*, *12*(3), 435–467.