Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

What Behaviour Cannot Tell Us About Belief: Obedience, Deception and Moral Injury

Manuel Lozano Rodríguez

perennidad@gmail.com

American University of Sovereign Nations

Research Article

Keywords: belief updating, identifiability, precision weighting, obedience, deception, moral injury

Posted Date: November 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11775353/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

When do compliance, deceptive reports and persistent distress warrant an inference about belief updating? This theoretical and methodological article uses analytical counterexamples and a synthetic measurement study. In a scalar Gaussian model, a change in judgment identifies the relative weight given to new evidence when the prior judgment and new input are known and differ. Different levels of uncertainty can produce the same change. A delta-rule learner can reproduce the same sequence under matched gains and a shared response model. Three applications distinguish action from endorsement in obedience, report selection from private learning in deception, and event appraisal from broader judgment in moral injury. A trust example separates two quantities: how far someone revises the probability of fair treatment by an original partner, and how much of that revision transfers to unfamiliar people. Measuring both judgments identifies these quantities under the assumed model. At equal measurement counts, crossing evidence credibility with target relations tests model restrictions while reducing precision in the focal condition. Yet a direct effect of the experience on expectations about strangers, bypassing the revised judgment about the original partner, can mimic stronger transfer in every condition. Independent calibration would estimate transfer from separate evidence capable of distinguishing these effects, fixing a prediction for the broader judgment. A discrepancy beyond measurement and calibration uncertainty would challenge transfer alone as calibrated.

From behaviour to belief

A person follows an order, repeats a false account or remains distressed after a moral transgression. What has been learned about that person's beliefs? The answer depends on how beliefs produce the observed response. Compliance may preserve income or avoid punishment. A report may serve an audience. Distress may concern a particular act or a judgment about one's entire life. Treating the response as a transparent record of belief removes the very problem requiring explanation.

The first example concerns uncertainty about one person. Suppose two people each put the chance that a trustee will treat them fairly at 80%. After receiving the same adverse information, both revise that chance to 35%. How sure is either person about the new figure? Both could have given the evidence the same weight relative to their earlier judgment while differing considerably in how uncertain they were about both. A rule that simply moves the estimate partway towards the new information can produce the same change. To distinguish the accounts, we need an observation on which they disagree.

Precision weighting gives one account of how evidence changes an estimate. In hierarchical Gaussian learning models, the relative precision of new information and prior estimates can govern updating, while a separate response model links estimates to decisions (Mathys et al., 2014). That link raises a question of identification: whether competing mechanisms make different observable predictions.

I examine this problem through analytical counterexamples, three distinct inference errors and a synthetic comparison of measurement designs. The examples and simulation data are constructed; the cited studies provide narrower empirical or conceptual anchors. The three cases put the same question

to different areas of psychology: what else must we observe before a response warrants an inference about belief? In obedience the missing evidence may concern the cost of refusing, in deception the audience for a report, and in moral injury the targets to which a judgment extends. The simulation then tests parameter recovery and model restrictions under specified noise and adverse observation processes.

An identification counterexample

Consider a scalar quantity x with a Gaussian prior of mean m and positive precision p. An observation y has independent, zero-mean Gaussian error with positive precision e. Precision is inverse variance. Multiplying the prior and likelihood and collecting the quadratic terms gives posterior precision p + e and posterior mean (pm + ey)/(p + e). Writing that mean as an update gives:

$$m_{post} = m + K(y - m), \quad K = \frac{e}{p + e}, \quad V_{post} = \frac{1}{p + e} \qquad (1)$$

When m and y are known and differ, the posterior mean determines K and therefore e/p. It cannot determine both absolute precisions. Multiplying p and e by the same positive constant leaves K and the mean unchanged while changing posterior variance. For example, p = 1 and e = 3 produce K = 0.75 and variance 0.25; p = 10 and e = 30 produce the same gain but variance 0.025. Mean updating therefore cannot distinguish these two uncertainty states. If y = m, the unchanged mean does not identify even the ratio.

A delta-rule learner changes its estimate d by a fraction α of the discrepancy. Nassar et al. (2010) derived an approximately Bayesian delta rule with a learning rate that varies in a changing environment. Here the comparison concerns an exact equality under matched gains:

$$d_{post} = d + \alpha\,(y - d) \qquad (2)$$

Setting d = m and α = K makes the updates identical for every y. Across trials, equality persists by induction if both learners begin at the same estimate, use matched gains at each history and carry their updated estimates forward in the same way. If their response models also assign the same probability to every observed response given those estimates and the task history, the resulting behavioural distributions, conditional on the same input sequence, are identical. More trials drawn from that same observation model cannot distinguish the matched descriptions.

This is a conditional equivalence, not evidence that all Bayesian and reinforcement-learning models are interchangeable. A freely history-dependent α may merely reparameterise the Bayesian rule; it is not automatically a simpler rival. A fixed-rate delta learner can differ from a Bayesian learner whose gain changes under specified uncertainty dynamics. Likewise, the pointwise scale invariance above need not survive arbitrary precision recursions across time. Claims about a longitudinal model require its actual transition rules.

The counterexample also concerns the observation model. If reports or choices depend separately on posterior uncertainty, matching the means alone will not guarantee identical behaviour. Yet confidence ratings do not automatically measure that uncertainty: reporting incentives, metacognitive calibration and response noise require their own specification. Parameter recovery within an assumed model and recovery of the correct model from competing generators are different checks (Wilson & Collins, 2019).

The result is limited but consequential. A good fit to mean updating can identify an effective gain under stated assumptions. It does not establish absolute precision, a uniquely Bayesian mechanism or its neural implementation. These are additional claims with additional evidential burdens.

Obedience and the inference of endorsement

Someone can follow an order to keep a job, avoid punishment or advance a purpose they share with the person who gave it. Milgram's (1963) experiment measured how far participants complied with instructions to administer what they believed were increasingly severe shocks. Kelman (1958) distinguished such public compliance from internalisation. Haslam and Reicher (2012) give identification with the authority and its cause a more active role in explaining obedience. In a military unit, loyalty to a political objective and discipline under command can lead to the same action (Lozano Rodríguez, 2025a). Observing the action leaves the degree of endorsement open.

Caspar et al. (2016) found that coercive instructions altered measures of experienced agency. Their result concerns the experience and processing of action under coercion; confidence in the commander's accuracy and moral responsibility require separate evidence. A study of obedience must therefore specify whether its endpoint concerns agency, endorsement or responsibility.

A discriminating study would vary learned source accuracy independently of sanctions while eliciting judgments under conditions that reduce the cost of disagreement. If compliance changes with sanctions while the separately elicited judgment remains stable, an action-policy explanation becomes plausible. If judgments track reversed source accuracy, reliability-sensitive learning receives support. Identifying a particular precision mechanism then requires observations that distinguish it from the matched alternatives in the Gaussian counterexample.

The relevant evidence concerns the decision and information available at the time. Reading an action as endorsement without considering the cost of refusal bypasses that requirement, as does inferring an earlier belief from disagreement reported only after the event.

Deception and the inference of private belief

A deceptive report is selected for another person to receive. It may therefore conceal successful learning. A witness can revise a private estimate after corrective evidence while continuing to state the earlier account because truthful reporting carries a cost. Repetition of the account then establishes persistence of a report, not persistence of the underlying belief.

In sender–receiver experiments, willingness to mislead varied with the consequences for sender and receiver (Gneezy, 2005). Kuran (1995) describes a related divergence in public life: people may voice a preference they privately reject when candour carries a cost. The same distinction between what someone says and what they have learned applies to deliberate misdirection in warfare (Lozano Rodríguez, 2025b). An experiment can vary the payoff for a report while holding the evidence fixed, then ask whether a separately elicited estimate moves with the report.

An audience can affect the judgment itself. Research on accountability finds that its effects depend on whom the person answers to and on the rules governing the exchange (Lerner & Tetlock, 1999). In online settings, the prospect of stigma can add a cost to disclosure (Lozano Rodríguez, 2022). Elicit private estimates and public reports separately, and vary evidence reliability independently of the costs of disclosure. A confidential estimate is still an elicited response. To distinguish self-deception from strategic reporting, look for biased acquisition or use of information as well.

Neural prediction does not remove this distinction. Lee et al. (2024) showed that predictors trained on deception could derive predictive power from processes present in a control task without deception. Successful classification and identification of the target process are separate achievements. A longer catalogue of activated regions would not repair that problem.

Moral injury and the inference of a mechanism

Shay (2014) locates one source of moral injury in betrayal by a legitimate authority in a high-stakes situation. The preliminary model of Litz et al. (2009) also considers perpetration, failure to prevent and witnessing. A study needs to record the exposure, how the person appraised it and which consequences persisted. That distinction matters when people who report the same event appraise it differently.

The consequences include changes in social connection and functioning (Litz & Walker, 2025). Conditions such as displacement and confinement can continue to constrain a person after the event (Lozano Rodríguez, 2023). A study can then ask whether distress continues alongside a changed appraisal while the threat itself is still present.

Research on fear generalisation distinguishes responses to a learned threat from responses to other cues (Dunsmoor & Paz, 2015). The examples below make a corresponding distinction between the objects of moral judgment. Recognising that one committed a wrongful act concerns a particular event; concluding that one is worthless in every relationship extends the judgment across targets. After betrayal, revising trust in a particular commander likewise differs from withdrawing trust from unrelated people. These analytical examples carry no individual diagnostic finding. They locate the distinction a measurement design must preserve: what changed about the event, and what transferred to other targets.

Persistent distress is compatible with several possibilities: an accurate appraisal of serious wrongdoing, an inaccurate attribution of responsibility, continued external threat or an overgeneralised negative

judgment. Even an observed learning difference would require a model specifying which belief changes and how that change produces the measured outcome. Distress alone cannot identify an excessively strong prior, excessive responsiveness to evidence or the level at which either occurs.

Useful observations would distinguish event-specific judgments from transfer to unrelated persons or situations and follow those judgments over time. Treatment selection and judgments of culpability require further evidence specific to those decisions. Where distress persists after an appraisal changes, the explanatory task is to account for that dissociation between belief change and the continuing outcome.

A worked comparison of updating and transfer

In a 2025 PsyArXiv preprint (https://doi.org/10.31234/osf.io/g327f), Herzog and McNally propose a belief-updating account of moral injury that relates responses to potentially morally injurious events to changes in moral beliefs. Their proposal motivates an experiment in which the credibility of event information varies. If the outcome is a broad change in belief, the experiment still needs to distinguish two quantities: what the participant learned about the event and how far that learning carried over.

A related identification problem arises in perception. Hahn et al. (2026) show how a prior and a loss function can be confounded in a Bayesian observer, and examine when measurements at multiple sensory noise levels can help recover model components. For trust, the question concerns two judgments: the chance of fair treatment by the original partner and the chance of fair treatment by unfamiliar people. Transfer means that a revision about the original partner changes the judgment about those other people. A direct effect changes that broader judgment through a route that bypasses the revision about the original partner. Measuring both judgments separates updating from apparent transfer under the proposed model; the analysis below shows how a proportional direct effect can produce the same answers.

There is an empirical reason to vary the relation between the original trustee and the new target. In two experiments, FeldmanHall et al. (2018) let participants learn about specific partners in a trust game and then choose among unfamiliar partners whose faces resembled those of the original players to varying degrees. Resemblance to a partner learned to be untrustworthy discouraged selection; resemblance to a trustworthy partner encouraged it. Relative to the neutral resemblance gradient, the reported similarity interactions for untrustworthy and trustworthy partners were − 0.65 and + 0.40 in Experiment 1, and − 0.49 and + 0.66 in Experiment 2 (Tables 1 and 2). Their differences, − 1.05 and − 1.15, are descriptive arithmetic contrasts calculated here by subtracting the trustworthy interaction from the untrustworthy interaction. These contrasts have no uncertainty estimates calculated here and express the opposing directions on the authors' coefficient scale.

These experiments show a change in partner choice. Paired probability judgments about both targets before and after the experience would measure the two changes in expectation. Interpreting their ratio as transfer requires the observation model and checks against competing explanations developed

below. Explaining choices through those judgments would then require a model linking what participants expect to whom they choose. In a separate series of 21 studies (N = 11,899), Roberts et al. (2026) reported that feedback about one partner changed both how likely participants thought others were to be trustworthy and what they expected trusting someone to feel like. The latter gives a rival route from experience to choice even when the belief about the next partner is held fixed. Together the findings motivate a design that varies the relation between targets, elicits the two probability judgments separately and tests how much those judgments explain about actual choices.

Take a constructed trust game that ends in a betrayal. Two people are being judged: the original trustee who broke trust and an unfamiliar trustee the participant has never met. For each, the participant states the probability of being treated fairly, once before the experience and once after. Prior information about a partner's moral character affects trust-game choices even when reinforcement rates are equal (Delgado et al., 2005). Eliciting a judgment about that partner gives the proposed study a belief measure to examine alongside choice.

Let mE be the participant's prior probability of fair treatment by the original trustee. Let y be the probability the event suggests to the participant: y = 0.20 means that, in the participant's reading, it points to a one-in-five chance of fair treatment. This is a subjective input to be elicited or otherwise justified. Write ΔE for the change in the judgment about the original trustee and ΔG for the change about unfamiliar trustees. Under pure transfer, g is the share of ΔE that reappears in ΔG, assumed here to lie between zero and one. For this example, assume linear transfer, no direct effect and an invariant, noiseless reporting map: the same belief always produces the same reported number. These assumptions give:

$$\Delta_E = K(y - m_E);\ \Delta_G = g\Delta_E = gK(y - m_E)$$

3

K is the effective gain from Eq. 1. The coefficient g specifies linear transfer across targets for this example. Eq. 3 is a proposed observation model whose application to moral injury would require evidence about how judgments transfer between people or situations.

Set both prior estimates to 0.80 and the event input to 0.20. With K = 0.75 and g = 1/3, the estimate for the particular trustee falls to 0.35, while the estimate for unfamiliar trustees falls to 0.65. With K = 0.25 and g = 1, the corresponding estimates are 0.65 and 0.65. The same 15-point decline in broader trust is one-third of a 45-point revision about the original trustee in the first case and the whole of a 15-point revision in the second. The updating gains differ threefold. A broad pre/post measure alone identifies the product gK, given the stipulated nonzero discrepancy, rather than either factor separately. Figure 1 shows where the two constructed cases diverge.

Under Eq. 3, with a known nonzero discrepancy, paired judgments give K = ΔE/(y − mE). When ΔE is also nonzero, they give g = ΔG/ΔE. At zero discrepancy the model predicts no change in either judgment for

any K or g. With a nonzero discrepancy and zero gain, the event judgment identifies K as zero while g remains unidentified. Zero transfer alone creates no such problem when the discrepancy and event change are nonzero. Small event changes make the transfer estimate sensitive to reporting noise. An experiment must therefore elicit the subjective event input or justify its measurement model, and test the effects of reporting incentives, scale differences and direct effects on recovery.

An extension of the proposed paradigm would elicit matched event-specific and unfamiliar-target judgments before and after the event, vary information credibility, and separately vary the supplied relation between the original and transfer targets. Reporting conditions should remain comparable. The algebra establishes a conditional identification requirement; selective effects of the manipulations and any connection between laboratory betrayal and clinical moral injury require separate empirical tests. Candidate models should specify whether each manipulation changes K, g or both, then undergo recovery checks under those alternatives.

An everyday version can make the measurement task concrete. Consider a constructed club example: before entrusting money to a member, officers record the chance that this member will account for it honestly and the corresponding chance for an unfamiliar member. The entrusted member later fails to account for the funds. One hypothetical officer lowers both judgments from 80% to 65% and says the incident itself suggests a 20% chance of honest accounting by the original member. Eq. 3 then gives K = 0.25 and an apparent g = 1. The two recorded judgments distinguish what changed about this member from what changed about a newcomer.

A study built around this example would elicit each pair privately, using the same questions and scale before and after the incident, and record the subsequent choice separately. If prior judgments were never recorded, the study would need prospective observations of another episode to calculate these changes. An independently justified calibration could then supply a prediction for the newcomer judgment. A discrepancy would prompt tests of a direct effect, a reporting change or the calibration's applicability. Choosing a safeguard such as two signatures requires additional evidence about losses, prevention and costs; the paired judgments contribute evidence about how the incident changed expectations. Assessing whether the change in expectations is warranted also requires evidence about the relationship between the two members.

Robustness of the gain–transfer distinction

The paired construction identifies K and g when the event discrepancy is known and nonzero, event updating is nonzero, and the observation map follows Eq. 3. Write D = y − mE for that discrepancy, keeping d for the delta-rule estimate in Eq. 2. A synthetic comparison examined broader judgments alone, paired event-specific and broader judgments, and a crossed design: paired judgments recorded at every combination of two credibility levels and two target relations. Each design received 48 scalar change measurements. This count provides a common observation allocation; a human study would also need to account for eliciting the subjective input and the work required of participants.

AI assistance. ChatGPT assisted with substantive redrafting, literature retrieval, mathematical checks, simulation programming, internal computational verification and document preparation. The author is responsible for the article and the supporting materials.

The broad-only design allocated all 48 measurements to one condition. The paired design allocated 24 to each judgment type. The crossed design used six measurements of each type in each of four conditions, with updating gains 0.25 and 0.75 and transfer coefficients 1/3 and 2/3. Its anchor combined gain 0.75 with transfer 1/3. All observations used independent Gaussian error on change measurements; sample means were generated with the corresponding standard errors. The archived configuration specifies nine scenarios, 10,000 replications per design and scenario, and seed 20261107. The resulting 270,000 simulated experiments concern the properties of the specified measurement model.

The scenarios varied noise, discrepancy strength, direct effects on the broader response, event-report scaling, the assumed event input and whether the target manipulation also changed event gain. At D = −0.6 and error standard deviation 0.08, paired judgments yielded median absolute errors of 0.0186 for K and 0.0258 for g. The corresponding crossed-design errors were 0.0265 and 0.0504. Dividing observations among conditions reduced anchor precision while permitting checks on the proposed selective influence of the manipulations. With a tenfold smaller discrepancy, median absolute transfer errors rose to 0.2654 and 0.5234, respectively. All estimates were retained, including those outside [0,1]. Table 1 compares the recovery errors for the two designs that record both judgments.

Table 1
Median absolute estimation error at the focal condition in the synthetic study

Quantity and scenario	Paired design	Crossed design
Updating gain (K), main	0.0186	0.0265
Transfer (g), main	0.0258	0.0504
Transfer (g), weak discrepancy	0.2654	0.5234

The event discrepancy was − 0.60 in the main scenario and − 0.06 in the weak-discrepancy scenario. The errors are medians of the absolute differences between the fitted and generating values. Crossing credibility with target relation also lets researchers check restrictions of the transfer model. The rejection rates below show how often the diagnostic flagged the simulated departures.

The diagnostic tested whether event means were equal across target relations and whether the two credibility rows were proportional in a weighted 2 × 3 matrix M. Let E(i,j) and G(i,j) be the event-specific and broader cell means, with i indexing credibility and j indexing target relation. Row i of M is [√2(E(i,1) + E(i,2))/2, G(i,1), G(i,2)]. Pooling the two event means and multiplying by √2 gives that entry the same error variance as each broader cell mean. The statistic adds the two squared event contrasts, each divided by two, to the squared smaller singular value of M, then divides by the known cell-mean error variance v = $\sigma^2/6$. The smaller singular value measures departure from proportional rows. Two event-equality

restrictions and two local rank-one restrictions give the four degrees of freedom of the asymptotic chi-square reference. The rule rejects above 9.4877. This approximation applies locally at nonzero signal with known variance; its false-alarm frequency near zero signal requires separate assessment.

The diagnostic can flag some direct effects. The constant-direct-effect scenario adds the same amount to every broader cell mean at a fixed discrepancy, whereas the proportional rival described next adds an amount that varies with the revision about the original partner. That proportional rival satisfies the diagnostic restrictions exactly. Under the main-noise generating model, the diagnostic rejected 5.12% of simulated experiments. Rejection reached 59.47% when target relation also changed event gain and 16.61% under the selected constant direct effect on broader judgment. Online Resource 1 gives the full statistic, scenarios and aggregate results.

A direct effect proportional to the revision about the original partner offers an exact competing explanation. If the broader response receives a direct contribution cKD, then $\Delta G = gKD + cKD = (g + c)KD$. The coefficient c stays fixed across the crossed conditions; the contribution cKD changes with the gain K and discrepancy D. Every condition is observationally equivalent to a pure-transfer model with $g' = g + c$ whenever the transformed coefficients are admissible and the joint noise distributions match. For example, c = 0.15 preserves admissibility for both transfer coefficients used here. Even a perfect fit then leaves transfer and the proportional direct effect unresolved. Figure 2 shows how the two accounts produce the same changes.

Independent calibration would estimate transfer from evidence capable of distinguishing it from the direct effect, and use that estimate to predict the broader judgment in the target study. Collecting a second dataset under the same ambiguous design would again identify their combined effect. The calibration needs a justified way to separate the two, uncertainty estimates and evidence that its value applies to the people and conditions being compared. With those requirements met, it supplies a prediction that a broader judgment can challenge.

Multiplying the event signal by a positive factor s before adding report noise gives the same joint distribution of reports as a model with $K' = sK$ and $g' = g/s$, provided the transformed parameters remain admissible and the joint error distribution is the same. The event-scaling rival poses a separate problem: it requires independent information about the reporting scale.

These results distinguish parameter recovery within a model from identification of the processes represented by its parameters. Broader judgments alone leave a many-to-one mapping from gain and transfer to observation. Paired judgments resolve that factorisation under the stipulated assumptions, while the crossed design examines additional restrictions. The exact adverse cases locate the further measurement evidence needed to interpret the recovered parameters as psychological mechanisms.

Evidence that could distinguish the accounts

The counterexample does not make behaviour scientifically useless. It identifies assumptions that an informative design must expose. Researchers should specify the latent quantity, its update rule, the transition between trials and the mapping to the recorded response. Named rivals such as threat, conflict or valuation become testable models only when those components generate distinct predictions under at least one feasible condition.

Parameter recovery asks whether known generating parameters can be estimated within a model. Model recovery asks whether data generated by each candidate can distinguish it from the alternatives. Both depend on the task and observation process; success on one does not establish the other. Held-out prediction tests performance on new observations. Models that make identical observable predictions remain indistinguishable on those observations (Wilson & Collins, 2019).

Return to the constructed participant whose two judgments fell from 80% to 65%, giving an apparent g = 1 under Eq. 3. Independent calibration would test a different assumption. Suppose separate evidence, able to distinguish transfer from the direct effect under comparable conditions, supports g = 1/3. A 15-point decline in the judgment about the original trustee would then predict a 5-point decline towards strangers, from 80% to 75%. The 65% value is the constructed outcome used for comparison. A real discrepancy of 10 points, beyond measurement and calibration uncertainty, would challenge transfer alone as calibrated. A direct effect would be one explanation to test alongside reporting differences and failure of the calibration to apply. Refitting g to the same broader change would absorb the discrepancy and restore the ambiguity. An independently justified value instead fixes the prediction: how far learning about one person should change expectations of others.

A further study could ask whether people reporting the same probability make different choices about buying more evidence. This proposal was not simulated here. Researchers would offer information at a stated cost, recording the choice to seek it as well as the initial judgment. For that choice to distinguish uncertainty states, each candidate must specify how uncertainty, expected benefit and cost affect the response. The decisive comparison is between their predictions at the same stated probability and cost.

Table 2 pairs each inference with the observation that would help distinguish it from a rival. The comparison establishes neither a common disorder nor a protocol that resolves all three. It assigns each application its own observation requirement and preserves those distinctions when interpreting the results.

Table 2
Distinct observation requirements across the three applications.

Application	Insufficient observation	Additional distinction required	Remaining limit
Obedience	Compliance	Sanctions versus separately elicited endorsement	Judgment patterns do not identify a unique precision mechanism
Deception	Persistence of a public account	Reporting incentives versus separately elicited estimates	An estimate remains a report
Moral injury	Persistent distress	Event-specific appraisal versus transfer to unrelated targets	Distress alone does not identify failed updating

Physiological observations can add discrimination when competing accounts predict different responses under controlled conditions. They cannot supply that discrimination merely by being physiological. The reverse-inference problem remains when a signal associated with several processes is used to identify one of them (Poldrack, 2006). A neural implementation requires predictions about physiology beyond the behavioural equivalence already demonstrated.

Conclusion

Two people can make the same downward revision in their estimate of fair treatment by a trustee while remaining differently uncertain about the new estimate. In the Gaussian model, the revision reveals the relative weight given to new evidence when the prior judgment and the differing event input are known. Distinguishing the remaining uncertainty requires an observation for which the competing accounts predict different responses.

The trust example adds a measurement requirement. Recording judgments about both the original partner and unfamiliar people separates the original revision from its apparent transfer under the specified model. Even across crossed conditions, a direct effect on expectations about strangers proportional to the original revision can mimic stronger transfer when its coefficient stays fixed, the implied transfer remains in range and reporting errors have matching distributions. Independent evidence capable of separating these effects would fix a prediction for the broader judgment. A discrepancy beyond measurement and calibration uncertainty would challenge transfer alone as calibrated. The next experiment should therefore record both judgments before and after the experience, elicit the participant's reading of the event, and establish an independently justified transfer estimate applicable to those participants and conditions.

Declarations

Funding: This research did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors.

Competing interests. None.

Ethics approval and consent to participate. Not applicable. The study used synthetic observations only.

Consent for publication. Not applicable.

Data availability. The data underlying the simulation results are available in Online Resource 1 (ESM_1.zip), which accompanies this article. The archive contains aggregate recovery and diagnostic results, saved diagnostic samples, and the configuration and code needed to regenerate the synthetic results.

Figure preparation. Figures 1 and 2 were drawn in Matplotlib from the worked example and the model equations.

References

1. Caspar, E. A., Christensen, J. F., Cleeremans, A., & Haggard, P. (2016). Coercion changes the sense of agency in the human brain. *Current Biology*, *26*(5), 585–592. https://doi.org/10.1016/j.cub.2015.12.067
2. Delgado, M. R., Frank, R. H., & Phelps, E. A. (2005). Perceptions of moral character modulate the neural systems of reward during the trust game. *Nature Neuroscience*, *8*(12), 1611–1618. https://doi.org/10.1038/nn1575
3. Dunsmoor, J. E., & Paz, R. (2015). Fear generalization and anxiety: Behavioral and neural mechanisms. *Biological Psychiatry*, *78*(5), 336–343. https://doi.org/10.1016/j.biopsych.2015.04.010
4. FeldmanHall, O., Dunsmoor, J. E., Tompary, A., Hunter, L. E., Todorov, A., & Phelps, E. A. (2018). Stimulus generalization as a mechanism for learning to trust. *Proceedings of the National Academy of Sciences, 115*(7), E1690–E1697. https://doi.org/10.1073/pnas.1715227115
5. Gneezy, U. (2005). Deception: The role of consequences. *American Economic Review*, *95*(1), 384–394. https://doi.org/10.1257/0002828053828662
6. Hahn, M., Wang, E., & Wei, X. X. (2026). Identifiability of Bayesian models of perception. *Proceedings of the National Academy of Sciences, 123*(37), e2601013123. https://doi.org/10.1073/pnas.2601013123
7. Haslam, S. A., & Reicher, S. D. (2012). Contesting the nature of conformity: What Milgram and Zimbardo's studies really show. *PLOS Biology*, *10*(11), e1001426. https://doi.org/10.1371/journal.pbio.1001426
8. Kelman, H. C. (1958). Compliance, identification, and internalization: Three processes of attitude change. *Journal of Conflict Resolution*, *2*(1), 51–60. https://doi.org/10.1177/002200275800200106
9. Kuran, T. (1995). *Private truths, public lies: The social consequences of preference falsification*. Harvard University Press. https://doi.org/10.2307/j.ctvt1sgqt

10. Lee, S., Niu, R., Zhu, L., Kayser, A. S., & Hsu, M. (2024). Distinguishing deception from its confounds by improving the validity of fMRI-based neural prediction. *Proceedings of the National Academy of Sciences, 121*(50), e2412881121. https://doi.org/10.1073/pnas.2412881121
11. Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin*, *125*(2), 255–275. https://doi.org/10.1037/0033-2909.125.2.255
12. Litz, B. T., Stein, N., Delaney, E., Lebowitz, L., Nash, W. P., Silva, C., & Maguen, S. (2009). Moral injury and moral repair in war veterans: A preliminary model and intervention strategy. *Clinical Psychology Review*, *29*(8), 695–706. https://doi.org/10.1016/j.cpr.2009.07.003
13. Litz, B. T., & Walker, H. E. (2025). Moral injury: An overview of conceptual, definitional, assessment, and treatment issues. *Annual Review of Clinical Psychology*, *21*, 251–277. https://doi.org/10.1146/annurev-clinpsy-081423-022604
14. Lozano Rodríguez, M. (2022). Shame on you! Digital ethnography, suicide, stigma, and warfare. In J. Chowdhury, H. Wahab, R. Saad, P. Roy, & J. Wronka (Eds.), *Practices, challenges, and prospects of digital ethnography as a multidisciplinary method* (pp. 198–222). IGI Global. https://doi.org/10.4018/978-1-6684-4190-9.ch014
15. Lozano Rodríguez, M. (2023). Bioethics of displacement and its implications. *IGI Global*. https://doi.org/10.4018/978-1-6684-4808-3
16. Lozano Rodríguez, M. (2025a). Political obedience as a military strategic asset: From Cyrus to the 2023 War in Gaza. *Military Strategy Magazine*, *10*(2), 27–33. https://doi.org/10.64148/msm.v10i2.3
17. Lozano Rodríguez, M. (2025b). Virtù and the War on Hamas: A Machiavellian reading of the strategies around the 2023 Israel-Hamas War. *Military Strategy Magazine*, *10*(1), 42–47. https://doi.org/10.64148/msm.v10i1.5
18. Mathys, C. D., Lomakina, E. I., Daunizeau, J., Iglesias, S., Brodersen, K. H., Friston, K. J., & Stephan, K. E. (2014). Uncertainty in perception and the Hierarchical Gaussian Filter. *Frontiers in Human Neuroscience*, *8*, 825. https://doi.org/10.3389/fnhum.2014.00825
19. Milgram, S. (1963). Behavioral study of obedience. *The Journal of Abnormal and Social Psychology*, *67*(4), 371–378. https://doi.org/10.1037/h0040525
20. Nassar, M. R., Wilson, R. C., Heasly, B., & Gold, J. I. (2010). An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *Journal of Neuroscience*, *30*(37), 12366–12378. https://doi.org/10.1523/JNEUROSCI.0822-10.2010
21. Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences*, *10*(2), 59–63. https://doi.org/10.1016/j.tics.2005.12.004
22. Roberts, A. R., Levine, E. E., & Risen, J. L. (2026). Learning to distrust: One trust experience changes the expected value of trust. *Journal of Experimental Social Psychology*, *125*, 104930. https://doi.org/10.1016/j.jesp.2026.104930
23. Shay, J. (2014). Moral injury. *Psychoanalytic Psychology*, *31*(2), 182–191. https://doi.org/10.1037/a0036090

24. Wilson, R. C., & Collins, A. G. E. (2019). Ten simple rules for the computational modeling of behavioral data. *eLife*, *8*, e49547. https://doi.org/10.7554/eLife.49547

Figures

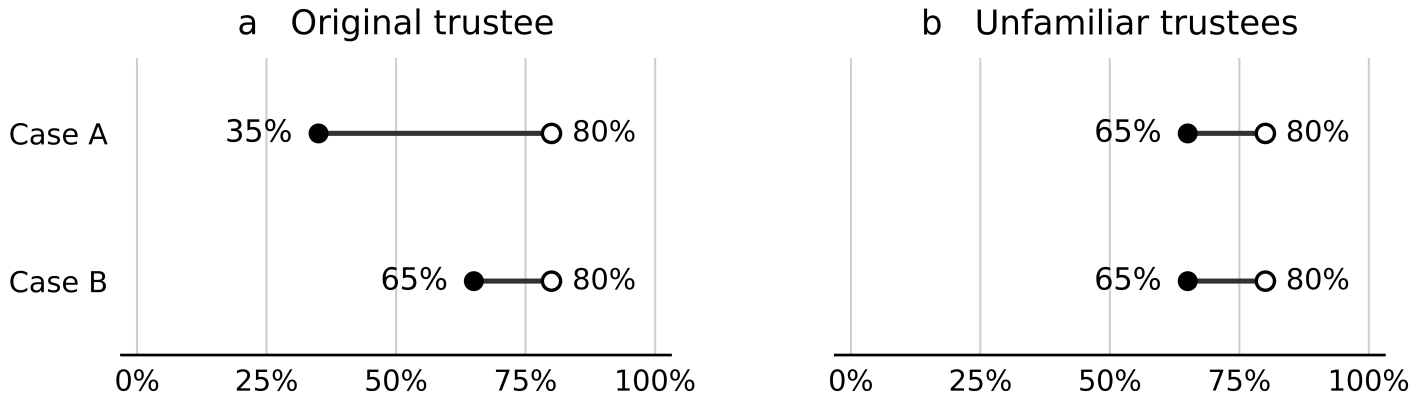

Figure 1

Estimated chance of fair treatment by the original trustee (a) and unfamiliar trustees (b) in two constructed cases under Equation 3. Open circles mark judgments before the event and filled circles mark judgments after it. Case A uses K = 0.75 and g = 1/3. Case B uses K = 0.25 and g = 1

Under Equation 3, with a known nonzero discrepancy, paired judgments give $K = \Delta E/(y - mE)$. When ΔE is also nonzero, they give $g = \Delta G/\Delta E$. At zero discrepancy the model predicts no change in either judgment for any K or g. With a nonzero discrepancy and zero gain, the event judgment identifies K as zero while g remains unidentified. Zero transfer alone creates no such problem when the discrepancy and event change are nonzero. Small event changes make the transfer estimate sensitive to reporting noise. An experiment must therefore elicit the subjective event input or justify its measurement model, and test the effects of reporting incentives, scale differences and direct effects on recovery.

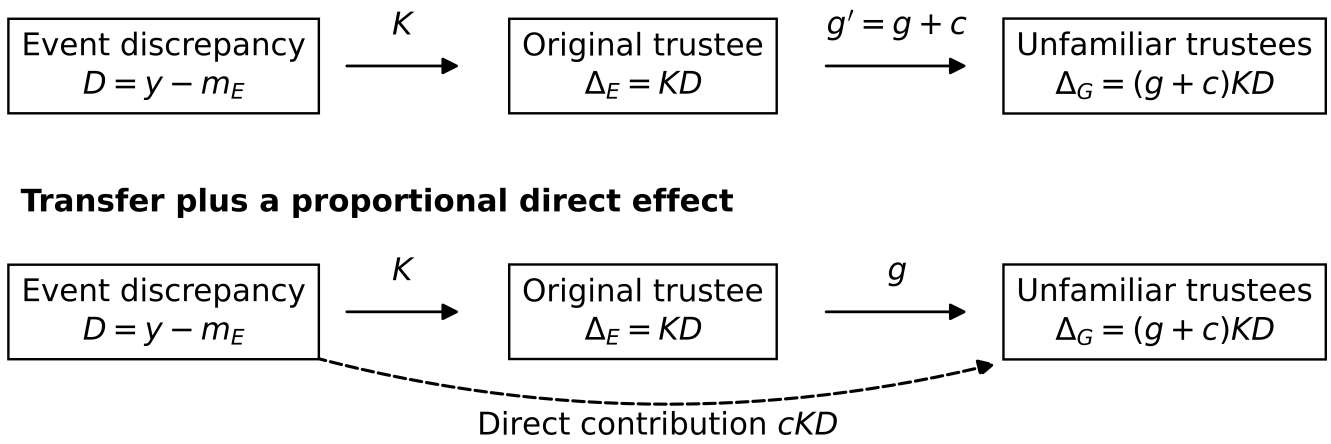

Figure 2

Pure transfer with an adjusted coefficient (a) and transfer plus a proportional direct effect (b). The dashed arrow adds cKD directly to the broader judgment. The two accounts have identical observable distributions across crossed conditions when c is fixed, g' is admissible and the joint reporting-error distributions match

Independent calibration would estimate transfer from evidence capable of distinguishing it from the direct effect, and use that estimate to predict the broader judgment in the target study. Collecting a second dataset under the same ambiguous design would again identify their combined effect. The calibration needs a justified way to separate the two, uncertainty estimates and evidence that its value applies to the people and conditions being compared. With those requirements met, it supplies a prediction that a broader judgment can challenge.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- IPBSResourcesWhatBehaviorCannotTell.zip
- OnlineResource1info.docx