COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 47 (2023) e13313

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13313

Categorical Updating in a Bayesian Propensity Problem

Stephen H. Dewitt,[a] Nine Adler,[a] Carmen Li,[a] Ekaterina Stoilova,[a] Norman E. Fenton,[b] David A. Lagnado[a]

[a]*Department of Experimental Psychology, University College London*
[b]*School of Electronic Engineering and Computer Science, Queen Mary University of London*

Received 4 August 2020; received in revised form 1 June 2023; accepted 20 June 2023

Abstract

We present three experiments using a novel problem in which participants update their estimates of propensities when faced with an uncertain new instance. We examine this using two different causal structures (common cause/common effect) and two different scenarios (agent-based/mechanical). In the first, participants must update their estimate of the propensity for two warring nations to successfully explode missiles after being told of a new explosion on the border between both nations. In the second, participants must update their estimate of the accuracy of two early warning tests for cancer when they produce conflicting reports about a patient. Across both experiments, we find two modal responses, representing around one-third of participants each. In the first, "Categorical" response, participants update propensity estimates as if they were certain about the single event, for example, certain that one of the nations was responsible for the latest explosion, or certain about which of the two tests is correct. In the second, "No change" response, participants make no update to their propensity estimates at all. Across the three experiments, the theory is developed and tested that these two responses in fact have a single representation of the problem: because the actual outcome is binary (only one of the nations could have launched the missile; the patient either has cancer or not), these participants believe it is incorrect to update propensities in a graded manner. They therefore operate on a "certainty threshold" basis, whereby, if they are certain enough about the single event, they will make the "Categorical" response, and if they are below this threshold, they will make the "No change" response. Ramifications are considered for the "categorical" response in particular, as this approach

Correspondence should be sent to Stephen H. Dewitt, Department of Experimental Psychology, University College London, 26 Bedford Way, London WC1H 0AP, UK. E-mail: dewitt.s.h@gmail.com

produces a positive-feedback dynamic similar to that seen in the belief polarization/confirmation bias literature.

Keywords: Propensity; Bayesian network; Uncertainty; Confirmation bias; Belief polarization; Causality

1. Introduction

1.1. Updating propensities

A “propensity” has been defined as a tendency for a system to behave in a particular way (e.g., Kahneman & Varey, 1990; Popper, 1959). Propensities are rife in everyday human experience (Tesic, Liefgreen, & Lagnado, 2020), and functioning in the world requires our ability to maintain and update accurate estimates of propensities (Keren & Teigen, 2001). We might need to maintain estimates of a difficult pupil’s propensity to misbehave on a daily basis, the propensity of a patient to relapse, or the propensity for a computer or machine to produce errors. In the simplest cases, a propensity can be estimated from a running tally of observed instances; however, it is often uncertain whether an instance has really occurred. If a teacher returns to a classroom to find something has been broken in their absence, is this another example of the pupil misbehaving, or did another student do it, knowing that the first would get the blame? A key question we are interested in is, in such uncertain circumstances, (1) how the particular instance is interpreted and (2) how propensity estimates are updated in light of the latest instance. If the child is assumed to be to blame based on their past misbehavior, are they then perceived as even more likely to misbehave in the future? Has another instance been added to their tally?

For example, consider the following problem which will be shown to participants in experiments 1 and 2: two nations, X and Y are independently testing new missile technology. Neither has any issues launching missiles, but their missiles do not always successfully explode. Each has launched six missiles: X’s launches have successfully exploded one time and Y’s launches have successfully exploded four times. You observe a new missile explosion on the border between X and Y but cannot determine the source. (1) Based only on the provided information, what is the probability that X or Y are the source of this latest missile explosion? (2) Further, what is your best estimate of the propensity for X’s and Y’s future missile launches to result in an explosion after this latest observation?

In answering question (1), individuals must represent the information pertaining to a single, indivisible event: X or Y must have been the source, and only one of them, but it is uncertain which of the two sources is responsible. Furthermore, the only evidence available to judge this is their differing historical propensity to produce successful explosions. In question (2), individuals must move from representing that information about a single, indivisible event to choosing how to use it in order to update propensities (Kahneman & Varey, 1990; Keren & Teigen, 2001; Tesic et al., 2020).

1.2. Bayesian model

The "propensity" question (2) is the main focus of this paper and one which we believe has not been studied in the judgment and decision-making literature previously, although a related kind of updating of a generic causal power/base rate has been investigated in causal learning (Cheng, 1997; Griffiths & Tenenbaum, 2005) and perception (Zylberberg, Wolpert, & Shadlen, 2018). The question asks the solver to update the propensity for each nation's future missile launches to explode (henceforth, "ME-propensity" [missile explosion propensity]) in light of the latest missile launch. This question has not been considered in classic Bayesian updating problems. For example, in the classic taxi-cab problem (e.g., Bar-Hillel, 1980), we are told of a hit-and-run car accident in a city, which we know was caused by one of two cab companies, the green and the blue company. We are told that the green company is more common in the city (85% vs. 15%); however, an eyewitness has claimed that the cab was blue, and under testing has been shown to have an 80% accuracy rate. In this and many similar problems, the question of interest has always been the probability of the uncertain event: what is the chance the cab is blue?

Our propensity question is instead equivalent to asking participants to update their estimate of the witness's accuracy in light of their claim that the cab is blue, or, in the similar medical diagnosis problem (Casscells, Schoenberger, & Graboys, 1978; Gigerenzer & Hoffrage, 1995), to update their estimate of the false-positive rate of the mammography machine. These figures have typically been assumed to be fixed (known with omniscient accuracy) in these classic problems in the field but of course, this is never the case in real-world reasoning, where our estimates of true parameters are always uncertain. Indeed, the taxicab problem does not in fact tell us how many trials were used to establish the witness's reliability or the testing conditions (see Birnbaum, 1983; Welsh & Navarro, 2012).

In answering the propensity question, our initial or prior ME-propensity for X and Y (before the latest explosion) is best estimated by the proportion of their historical launches which have exploded. Using a classical approach would produce estimates of 16.7% (1/6) for X and 66.7% (4/6) for Y. However, given that the purpose of our model is to observe updates in these figures, the classical approach has a crucial limitation. For figures of either 0/6 or 6/6 (which for example can be seen in experiment 3), the classical approach produces estimates of 0% or 100% with no variance and which, therefore, cannot be updated. A long-known alternative is provided by Laplace's (1814) "rule of succession," which adds 1 to the numerator, and 2 to the denominator, as can be seen in Eq. 1. This approach assumes that, before making any observations, we have observed one success, and one failure (1/2), and is commonly used in situations with small numbers of observations as it produces less extreme outputs than the classical approach. However, it also converges on the classical outputs as *n* increases.

Eq. 1. Calculation of prior ME-propensities for X and Y using Laplace's (1814) "rule of succession"

$$P\left(E|X_f\right) = \frac{s+1}{n+2} = \frac{1+1}{6+2} = \frac{2}{8} = 0.25$$

$$P\left(E|Y_f\right) = \frac{s+1}{n+2} = \frac{4+1}{6+2} = \frac{5}{8} = 0.625$$

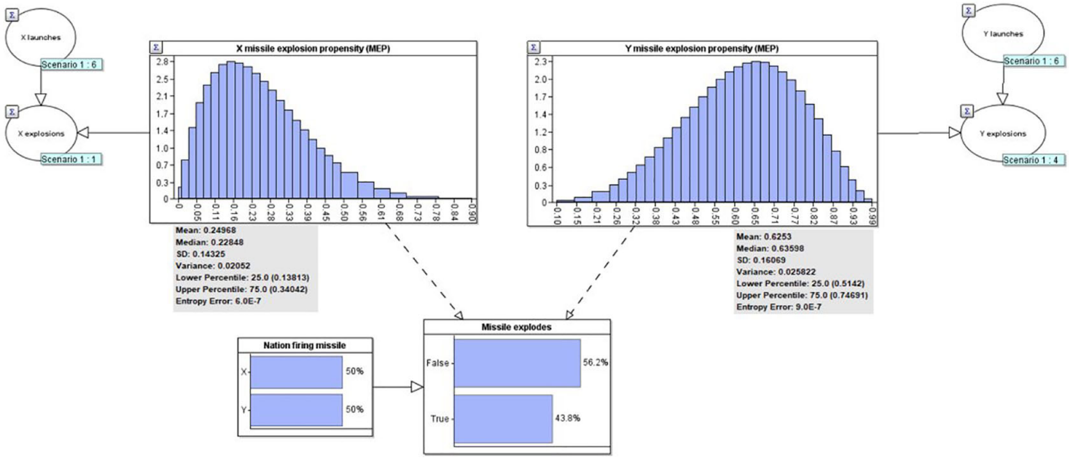


Fig. 1. A BN model depicting the missile scenario prior to the observation of the latest explosion.

Laplacian estimates also produce outputs in line with a Bayesian estimate updated from a uniform prior (where all outcomes are initially equally likely), which is how we implement this approach. We constructed a Bayesian network (BN: Fig. 1) to model this question in the present scenario. While we hope this model is of interest to many readers and adds depth to understanding the experimental data, it is important to note that a full understanding of this is not necessary to interpret our participants' responses to the missile problem. A BN is a directed graph whose nodes represent uncertain variables, and where an arc (or arrow) between two nodes depicts a causal or influential relationship (see Fenton & Neil [2018] for full details of BN's). In addition to the graph structure, each node has an associated probability table which defines the prior probability distribution for the associated variable, conditioned (where a node has parents) on its parent variables. When the state of a node is observed (e.g., the latest missile explodes), the known value is entered into the BN and a propagation algorithm updates the probability distributions for all unobserved nodes.

The model in Fig. 1 depicts the situation before observing the explosion, but with the information about previous missile launches. In the two upper distributions (e.g., "X Missile Explosion propensity [ME-propensity]"), the probability distribution for the next missile launch from X and Y exploding can be seen. These are uniform (0,1) distributions (all outcomes are initially equally likely) updated based upon the two circular nodes adjacent to them, depicting the number of previous successful explosions and total attempts for each nation (1/6 for X far left; 4/6 for Y far right). In the model, the number of successful explosions is defined as having a Binomial (n,p) distribution where n is the number of attempts and p is the propensity to explode. Given the observed values, the uniform priors for ME-propensity are updated into Beta distributions whose respective means are 25.0% for X and 62.5% for Y. These are identical to those calculated using the Laplace equation (Eq. 1) but deviate from the classical means of 16.7% for X and 66.6% for Y. More details on the model are provided at the online repository (https://osf.io/etsav/?view_only=c58c6300dca24be7859cbcb1cfceb709).

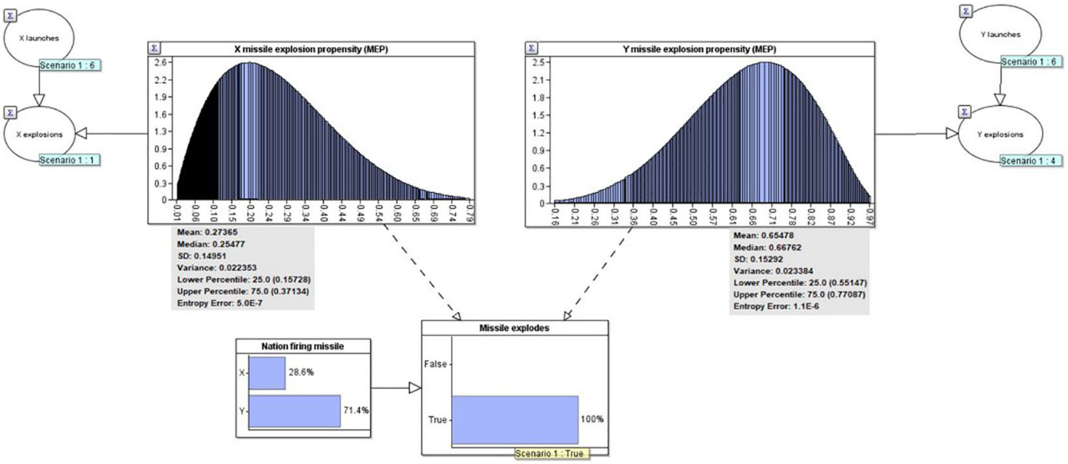


Fig. 2. A BN model depicting the missile scenario after the observation of the latest explosion.

Table 1
Bayesian model outputs for initial, final, and change in mean ME-propensity estimates for X and Y across a range of initial frequencies

Prior frequencies	X ME-propensity Initial → Final (Change)	Y ME-propensity Initial → Final (Change)
X [1/6] Y [4/6]	25.0 → 27.4 (**+2.4**)	62.5 → 65.5 (**+3.0**)
X [2/6] Y [4/6]	37.5 → 40.1 (**+2.6**)	62.5 → 65.1 (**+2.6**)
X [2/6] Y [5/6]	37.5 → 39.8 (**+2.3**)	74.9 → 76.9 (**+2.0**)
X [1/6] Y [5/6]	25.0 → 27.1 (**+2.1**)	75.0 → 77.1 (**+2.1**)

The two propensity distributions are combined into a single Boolean variable “Missile explodes” at the bottom, which gives the probability that the next missile launched will explode given that we do not know whether X or Y will launch it. This is also fed into by a Boolean variable (two possible outcomes) “Nation firing missile” which provides the probability that each nation will launch the next missile, which we assume to be 50:50. In this case, therefore, with 50:50 weighting, the “true” probability of the “Missile explodes” node is simply the average ME-propensity for X and Y ((25.0 + 62.5) /2 = 43.8).

Upon observing that the missile has exploded (achieved by “observing” the “Missile explodes” node as “True”), the BN (Fig. 2) automatically calculates the revised ME-propensity means to be 27.4% for X (a 2.4% absolute increase from 25.0%) and 65.5% for Y (a 3.0% absolute increase from 62.5%). We can see that it also updates its estimate of who launched the latest missile from 50:50 to 71.4% likely to be Y (middle left, “Nation firing missile” node).

The actual degree of mean ME-propensity increase for X and Y depends not only on their prior ME-propensity, but on the prior ME-propensity of the other nation also. In Table 1, we

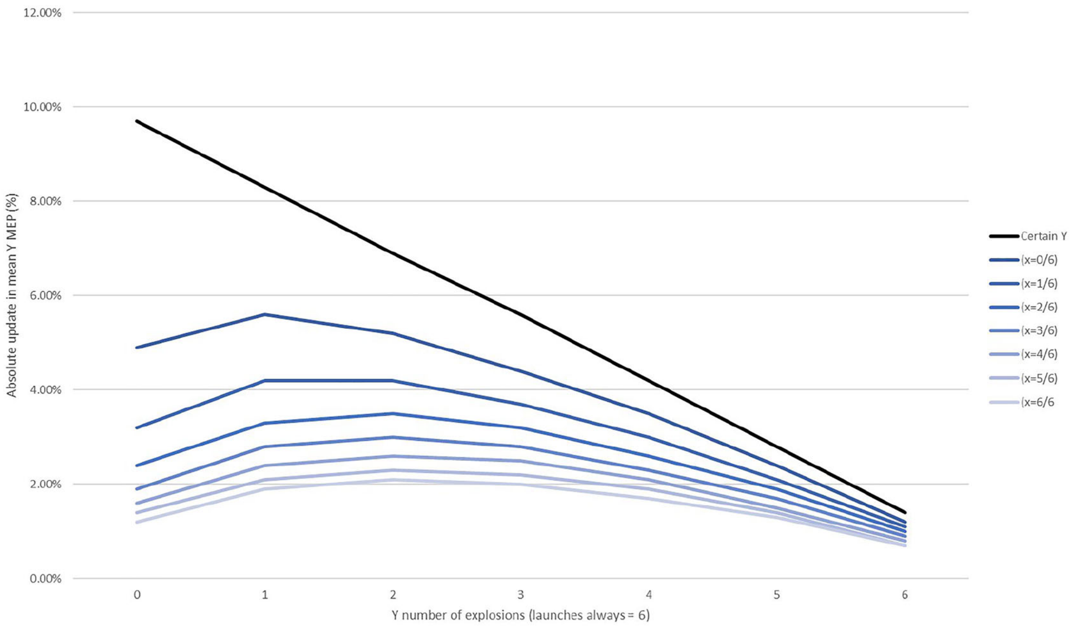


Fig. 3. A depiction of the absolute mean ME-propensity increase in Y (vertical axis) following the latest missile explosion for a range of Y prior frequencies (horizontal axis) and a range of X prior frequencies (different color lines). The increase when we are certain that Y launched the missile is also shown in black.

can see the mean ME-propensity updates that the model produces for a range of simulated prior mean ME-propensity values for X and Y before versus after the latest explosion.

In the first row, with the figures used in the problem, we can see a mean ME-propensity increase of 2.4% for X and 3.0% for Y. In the second row, we increase the prior frequencies for X to 2/6 but keep Y at 4/6. The subsequent mean ME-propensity update for X is larger (2.6%), while the update for Y is smaller (2.6%) compared to row 1. While the prior for Y has not changed, the smaller increase for Y (and the larger increase for X) here is because X is now more likely to be the source of the missile so takes a larger share of the update. If we now move to row 3, where we increase the prior frequencies for Y to 5/6 but keep X at 2/6, we see a mean ME-propensity update for X of 2.3% and 2.0% for Y. The increase for X is smaller compared to row 2 because Y is now more likely to be the source of the latest missile. However, note that the increase for Y is also smaller compared to row 2, despite being more likely to be the source. This is because of a final effect of diminishing increases as one approaches a mean ME-propensity of 100%. The actual increase, therefore, depends on these two factors: one's own prior ME-propensity (the closer to 100%, the smaller the increase) and the probability that they launched the missile (which is itself dependent on both their own prior ME-propensity, and the prior ME-propensity of the nation). The outcome of both of these dynamics can be seen in Fig. 3. Here, we show the percentage update in mean ME-propensity for Y (vertical axis) for a range of prior frequencies for Y (from 0/6 to 6/6: horizontal axis), and also for a range of prior frequencies for X

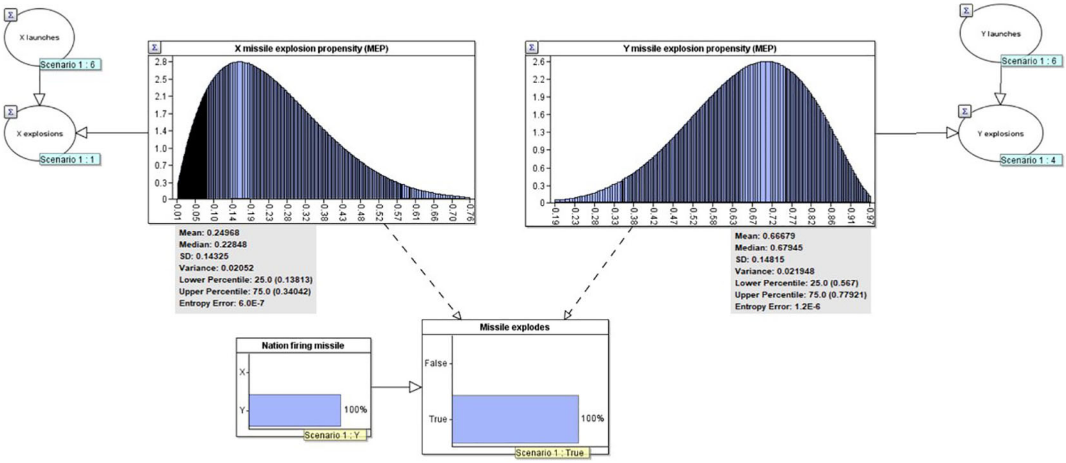


Fig. 4. A Bayesian network depicting the missiles scenario where we have observed the missile explode, and we also know that it came from Y.

(from 0/6 to 6/6: blue lines). The update value when we are certain Y launched the missile (see Fig. 4 for the BN of this) can also be seen in black. Looking at the steadily decreasing black line, when we are certain Y launched the missile, we can most clearly see the diminishing increase in mean ME-propensity as Y prior mean ME-propensity approaches 100% (i.e., from 0/6 to 6/6). In the blue lines, we can first see the positive effect of Y prior mean ME-propensity on the increase seen: these two competing dynamics result in a peak for each of the blue lines around a Y prior mean ME-propensity of 1/6 or 2/6. Finally, in the difference between the blue lines, we can also see the effect of the prior mean ME-propensity of X. Y mean ME-propensity increases more when X prior mean ME-propensity is lower.

1.3. Categorical updating

While the time we have taken to dig into the dynamics of the model will be valuable for the remainder of the paper, the key outcome for present purposes is that the mean ME-propensity for both X and Y increase following the explosion (sometimes equally, sometimes not, depending on the particular frequencies used). In an earlier exploratory study of updating propensities (Dewitt, Fenton, Liefgreen, & Lagnado, 2020), we told participants in a modified version of the taxi cab problem that the witness had been tested five times, correctly identifying the cab color four times (rather than just providing this as a fixed 80% as in the original) and also requested them to update their estimate of the witness's accuracy after they report the cab is blue. While this should go down, because their report is in opposition to the only other evidence that 85% of cabs are green, around a quarter of our participants in fact increased their estimate of the witness's accuracy. We collected qualitative data from these participants, explaining their reasoning, and this hinted that they may have been assuming that the witness was correct when updating their estimate of their accuracy (effectively increasing their tally from 4/5 to 5/6). However, the structure of that problem was not capable of definitively demonstrating that this response was based on this "categorical" or assumption-based reasoning, because there was only one propensity to update (the witness's accuracy when identifying cab colors), so "graded" updating could not be ruled out.

In the first two experiments in this paper, this kind of assumption-based or categorical updating can be definitively linked to a particular pattern of ME-propensity updating responses. If participants act upon the assumption that Y is responsible for the latest explosion, their responses should be equivalent to the pattern of updating seen in Fig. 4. Here, we have also "observed" on the "Nation Firing Missile" node that we know that Y launched the missile. In this case, the mean ME-propensity for X returns to its original level ($\sim$25.0%), while the mean ME-propensity for Y increases to 66.7%, which is equivalent to a Laplacian estimate for 5/7 (i.e., 6/9 = 0.666…), that is, Y has now launched one more successful explosion.

There is substantial literature to support the possibility of categorical reasoning on a probabilistic problem like this. Human reasoners are known to deal with and update probabilities in a categorical manner in a range of circumstances (e.g., Chen, Ross, & Murphy, 2014; Gallistel, Krishan, Liu, Miller, & Latham, 2014; Murphy & Ross, 1994; Sanborn, Noguchi, Tripp, & Stewart, 2020; Yu, Dayan, & Cohen, 2009). Indeed, "black and white" thinking, where we become convinced that a particular individual or group is responsible for an outcome, even when it is technically uncertain, is very common in social reasoning (e.g., Dyson, 2009; Frenkel-Brunswik, 1949; Glad, 1983) and could be a candidate process for how such categorical updating could come about.

If confirmed, this response would be of considerable interest, as the use of categorical updating in this problem would produce an interesting dynamic: Y would be assumed to be the source of the missile based solely on their historical propensity, and then that propensity would be updated based on that assumption. This reasoning has a circular characteristic similar to that seen in the belief polarization/confirmation bias literature (Cook & Lewandowsky, 2016; Fryer, Harms, & Jackson, 2015; Jern, Chang, & Kemp, 2014; Lord, Ross, & Lepper, 1979; Nickerson, 1998; Plous, 1991; Rabin & Shrag, 1999) and could be a mechanism by which initial beliefs about other agents (e.g., a belief that a pupil is badly behaved) could become self-reinforcing.

Therefore, if participants do update ME-propensity estimates based on the assumption that Y launched the missile, they should increase their ME-propensity estimate for Y only, leaving X unchanged, or perhaps reduced. In the first experiment in this paper, we primarily seek to determine if participants do indeed update ME-propensity estimates consistent with this categorical approach. Further, at this early stage of the investigation, we seek to explore the reasoning processes of our participants by asking them to explain their reasoning in an open text box.

2. Experiment 1

Beyond the simple exploration of participant responses and thought processes, we sought to include an experimental test of the tentative "black and white thinking" theory of why

"Y only" responses may occur, by introducing four conditions. These conditions only varied in the way in which participants were asked about whether X or Y were responsible for the latest missile launch (the single event question). In the control condition, they were not asked about this at all. In the "categorical" condition, they were asked to choose from binary options whether X or Y were responsible, and in two different types of probability conditions (explained below), they were asked to give a graded/probabilistic estimate of this. We intended these conditions to either encourage black and white thinking (categorical condition) or discourage it (the two probabilistic conditions). If black and white thinking takes place in this problem, we would expect more "Y only" responses in the categorical condition than the control condition, and more in the control condition than either of the probabilistic conditions.

2.1. Method

2.1.1. Participants

Two hundred and fifty-five participants for experiment 1 were recruited from Amazon MTurk and were required to have completed at least 50 tasks on the platform previously, have a 90% approval rating, and be based in the United States. One hundred and twenty-five (49.0%) self-identified as female, 125 (49.0%) as male, and 5 (2.0%) as "other." Mean age was 37.9 (SD = 11.8) with a minimum of 18 and a maximum of 72. Participants were asked their highest education level, and 32.9% reported high school, 47.8% bachelor's degree, 13.3% master's degree, 1.6% doctoral degree, and 4.3% other. Participants were also asked their profession and 43.9% reported "professional/managerial," 30.6% "labor/service," 4.7% "student," 10.6% "unemployed," and 10.2% "other."

2.1.2. Design

Participants all had the same experience, other than that the question participants were asked about who launched the missile (the "single event" question) was manipulated across four conditions: the control condition (participants were asked nothing); the "categorical" condition (participants were asked to choose categorically between the two nations) and two probabilistic conditions using slightly different formats. In the first ("Probability separate"), they were asked to provide probabilities on separate percentage scales for each nation, while in the second ("Probability combined"), they were asked to represent who was more likely to have launched the missile on a single scale from "Definitely X," through neutral, to "Definitely Y"). While the former was intended to allow for a direct measure of probability (e.g., participants could choose "73%"), the latter was intended to allow for a more fuzzy, intuitive representation (participants just chose a point on the scale which had no other labels than those three major ones).

2.1.3. Procedure and materials

All participants were presented with the background information (see publicly available materials and data at the online repository: https://osf.io/etsav/?view_only=c58c6300dca24be7859cbcb1cfceb709) including the 1/6 missile successes for X (in the experiment, X was named "Oclar" and Y "Trubia" to increase salience) and 4/6 for Y. They were then asked to provide an estimate for each nation's ME-propensity in percentage form on a sliding scale (one for each nation). All participants were then presented with information regarding the missile explosion of an unknown source. At this point, participants in the different conditions were provided the questions about who launched the missile detailed above.

Following this, all participants were asked, in light of the latest explosion, to indicate whether they would like to alter their estimate of each nation's ME-propensity. To indicate this, they were provided with two scales, one for each nation, with 7 points (Increase a lot; Increase some; Increase a little; Make no change; Decrease a little; Decrease some; Decrease a lot). We used this "change" method, rather than requesting a posterior point estimate and calculating change in comparison to the prior for several reasons. First, sliders can be noisy (individuals can struggle to get 100-point sliders exactly where they want) and participants may forget their original estimate. For these reasons, previous experiments have varied in how much "change" is considered to be zero (e.g., within 1%, 2%, etc.). Capturing zero change definitively was very important in this experiment, so we wanted to give individuals a clear opportunity to make that choice. Further, we thought that a percentage approach may encourage individuals to make the meta-assumption that a mathematical approach was expected, which we wished to avoid. We sought to capture intuitive responses to the problem and were more interested in direction (i.e., increase, no change, and decrease) than precise magnitude. After each of the above estimates, participants were requested to record their thought processes via open text boxes.

2.2. *Results*

2.2.1. *Quantitative data*

The responses to the posterior question were coded into response types based on the combined pattern of both their X and Y ME-propensity updates. As can be seen in Fig. 5, out of 255 individuals, the most common response ("Y only") was to increase the ME-propensity estimate for Y, and either make no change to, or reduce their ME-propensity estimate for X ($N = 88$, 34.5%) and the second was to make no change to either ME-propensity estimate ($N = 80$, 31.4%). Some individuals increased their ME-propensity estimate for both nations, but Y more ($N = 22$, 8.6%: labeled "Both, Y more") and a final subset increased both ME-propensities an equal amount ($N = 28$, 11.0%: labeled "Both equal"). Thirty-seven (14.5%) individuals provided a range of other responses.

The manipulation across four conditions (which manipulated the question individuals were asked about who launched the missile) was designed to reduce "black and white" thinking and ultimately, the "Y only" response. However, "Y only" response proportions were similar across conditions, as can be seen in Table 2. A binary logistic regression was run with condition as a factor predictor variable, and "Y only" as a binary criterion variable. No overall effect of condition was seen (X^2 (3) = 1.19, p = .755). We also tested the "linear" expectation that the categorical condition would produce the most "Y only" responses, followed by the control condition, followed by both of the probabilistic conditions combined. For this purpose, we created a "Conditions combined" variable, where "categorical" was coded "1," control "2," and both probabilistic conditions coded as "3." We used this as a predictor vari-

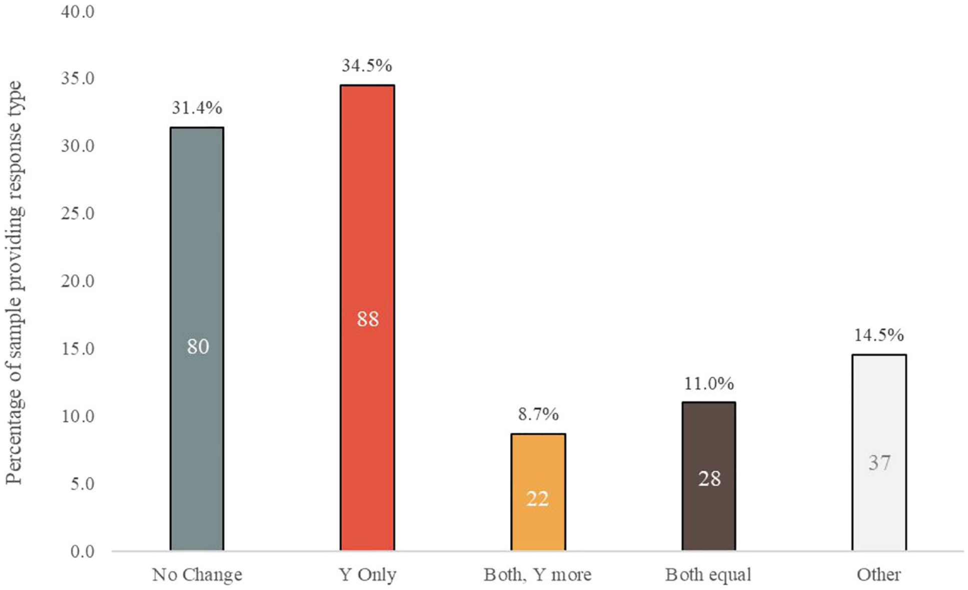


Fig. 5. The percentage of participants providing each of the response types.

Table 2
The proportion of participants giving the "Y only" response across all four conditions in experiment 1

Condition	Proportion giving "Y only" response
1 (Categorical)	39.4%
2 (Control)	34.7%
3 (Probabilistic combined)	32.8%
4 (Probabilistic separate)	30.5%

able in a binary logistic regression model with a binary criterion variable called "Y only" which was coded with a "1" if participants gave that response, and a "0" otherwise (X^2 (1) = 1.11, $p = .292$).

Furthermore, among "Y only" responders within the "Probabilistic separate" condition (the only condition to request a percentage estimate that X/Y launched the missile), almost all individuals (bar one) provided a non-100% assessment of who launched the missile, suggesting that they were not in fact certain that Y had launched the latest missile. The mean response from "Y only" responders was 77.7% for Y and 22.1% for X. In comparison, for "No change" responders, the mean estimate that Y launched the missile was 68.1% and for X, 29.6%, suggesting that certainty may be a factor underlying this response type. Within the "Probabilistic separate" condition, an independent samples *t*-test was run to specifically compare "Y only" and "No change" response types on their probability assessments that Y launched the missile ($t[36] = 2.4$, $p = .022$).

Table 3
The final agreed coding proportions for the five main codes, as well as "Unclassified/other" for both "No change" and "Y only" response types

	n	Historical propensities	Probably Y	Don't know who launched it	More data/ information needed	Nothing has changed	Unclassified/ other
No change	80	16.3%	6.3%	47.5%	10.0%	16.3%	26.3%
Y only	88	35.2%	62.5%	5.7%	0.0%	0.0%	21.6%
Inter-rater *R*		83.3%	89.9%	91.7%	95.8%	76.8%	81.5%

Note. Initial inter-rater reliability for each code can also be seen in the bottom row.

2.2.2. Qualitative data

Open text boxes were provided for responders to explain their posterior ME-propensity updates for both X and Y. "Y only" and "No change" responses were first-coded by the first author and second-coded by the third author, who had no previous involvement in the experiment. The first coder familiarized themselves with the data and developed the coding scheme seen in Table 3 of the five main codes plus a "bin" code of "unclassified/other" for responses which either did not fit in the main five codes, or where the thought process could not be confidently determined. The first coder sent written instructions to the second coder, which can be seen in the online repository. Both coders then separately assigned the codes that they thought applicable to each response. This was undertaken blind to all aspects of the response other than the open text data, including condition and response type (i.e., the coders could not see whether the response was "Y only" or "No change"). Responses could be assigned any number of the main five codes (they were not mutually exclusive); however, a response was only coded "unclassified/other" if it could not be assigned one of the main five codes. Initial inter-rater reliability for each code can be seen in the bottom row of Table 3. Following this, the first and second coders met in a discussion session to resolve discrepancies. First, taking a conservative approach, if either of the two coders had assigned the "unclassified/other" code, this was applied as the final code, without discussion. This approach should reduce the degree to which the proportions of the five main codes reflect the coders' subjective interpretation. The remaining discrepancies were resolved through discussion. In the online repository, the initial codes assigned by each coder, as well as highlighted discrepancies, and the final code decision can be seen.

2.2.3. Historical propensities

This code was assigned any time that a participant referred to the prior success of X or Y, either numerically, for example, P137 "they have 4 of 6 successful attempts the numbers speak for themselves" or non-numerically, for example, P203 "Y has a history of more successful attempts." This was more common among Y only responders (35.2%) and often co-occurred with the "Probably Y" code below, for example, "Based on past performance I believe it was Y that launched it."

2.2.4. *Probably Y*

The most common code among "Y only" responders (62.5%) was "Probably Y" wherein they stated their belief that Y had launched the latest missile (or that they believed X had not launched the missile). Only in a few cases was this expressed with certainty, however, with all others using uncertain language. For example, P34 said "I assumed that Y made the successful explosion this time." Applying a simple content analysis, out of those 55, 13 used the word "believe," 13 used "likely" or "unlikely," nine used "think," eight used some form of "assume," six used some form of "probable," and three used the word "guess" (others included "doubt" [that X launched it], "feel," and "figure"). This simple content analysis can also be seen in the same data set in the online repository. To give further examples, P43 stated "Since I am operating on the assumption that Y launched the successful missile, X is irrelevant – their proficiency rating doesn't change" and P39 stated, "I think they [Y] are the ones who launched and exploded the missile, so my [propensity] rating of them would go up slightly." Finally, four participants clearly stated that they believed that Y now had 5/7 successful attempts, for example, P186 said "A 5th success out of 7 raises their success rate from 66 to 71."

2.2.5. *Don't know who launched it*

The most common code among "No change" responders (47.5%) was "Don't know who launched it." To be assigned this code, it was not simply enough to express non-100% certainty that Y launched the missile, as that would also include everyone assigned the "Probably Y" code. Instead, this had to involve an explicit statement that they were not sure who launched the latest missile. For example, P13 said "I would have a hunch but would not make an assessment until I had solid knowledge of who launched the missile," P212 said "Since we don't know who it was, we can't change our estimate," P158 said "I don't want to make a change until I know for sure," and P180 said "We don't know where the missile came from so my impression of them doesn't change."

2.2.6. *"More data/information needed' and "Nothing has changed"*

These were the least common codes and were only seen among "No change" responders and overlapped considerably with "Don't know who launched it." For example, for "More data/information needed," P191 said "there is insufficient evidence to change the estimate" and P163 said "Not enough info to change." For "Nothing has changed," P243 said "I see no reason for any changes" and P233 said "I would make no change as I have no new facts." These individuals could be expressing the same sentiment as those coded "Don't know who launched it" above, but this requires some reading between the lines, so we used separate codes.

2.2.7. *"Unclassified/other"*

It is important to note that these responses were not primarily "junk/spam" responses, but either represented unique or highly uncommon thought processes, or simply did not convey a clear reason, for example, P7 who said "I don't think this missile launch will greatly change how I look at Y" and P192 who said "I'm making no changes due to the circumstances."

2.3. Discussion

This experiment primarily aimed to explore participant responses to this novel problem. We found a majority of individuals giving the "Y only" response, where only the ME-propensity estimate for Y was increased, with X either unchanged or reduced. This response implicitly ascribes the entire event to Y (it is the correct response if one were certain Y was responsible). It was expected to occur, and it was considered possible that "black and white thinking" in terms of who launched the missile may have been responsible. However, our results undermine this as a suitable explanation. First, inferential tests suggest that the manipulation based on this theory did not affect the proportion of individuals providing this response (although it should be noted that the proportions do go in the expected direction). Furthermore, all "Y only" responders (bar one) in fact did represent the single event probabilistically in the "probabilistic separate" condition, when asked what the chance of Y/X launching the missile was (Y mean: 77.7%; X mean: 22.1%). It was only when they came to update the propensities for each nation, that these individuals showed any form of categorical representation of the problem.

This suggests that "black and white thinking" is not a strong candidate for explaining our participants' responses. It appears individuals are willing/able to express as a probability whether Y was the source of the missile but appear unwilling/unable to "use" that information in a probabilistic manner when updating propensities. The open text responses to propensity updates provide further evidence of this point. Out of 88 "Y only" responders, only a few individuals expressed certainty that Y had launched the missile. The majority stated as their reason that Y was more likely to have launched the latest missile. However, no individual explained why they did not "split" the event in a graded manner between the two nations (with more going to Y, i.e., the "Both, Y more" response), suggesting that they may not have considered this as an option. It is possible, therefore, that these individuals, unlike the "Both, Y more" responders, may believe that the event cannot be split between X and Y, and must be ascribed in full to one or the other. This belief may be due to the fact that of course, in reality, either X or Y must have launched the missile. However, a subjective Bayesian approach allows us to update our propensity estimates based upon our uncertainty about a binary event.

Another candidate explanation for the cognitive process underlying the "Y only" response comes from another set of literature on human reasoning during multistage inferences. Several authors (e.g., Gettys, Kelly, & Peterson, 1973; Johnson, Merchant, & Keil, 2020) have shown that under multistage probabilistic inferences, human reasoners tend to treat intermediary probabilities as certainties when used in later calculations. Gettys et al. (1973) define a multistage inference as:

> "…a series of single-stage inferences where the output of each previous stage becomes the input to the next stage" (Gettys et al., 1973 pp. 364)

Johnson, Merchant, and Keil (2020) gave the following example:

> "Suppose you are unsure whether the Fed will raise interest rates. Depending on this decision, congress may attempt fiscal stimulus; depending on Congress's decision, the CEO of Citigroup may decrease capital reserves; and depending on the CEO's decision, SEC regulators may tighten enforcement of certain rules" (Johnson et al., 2020 pp. 1430).

Perhaps to avoid combinatorial explosion and reduce computational demand, under such circumstances, both sets of authors have proposed that intermediary probabilities may be converted into certainties. For example, if it is very likely that Citigroup would decrease capital reserves under the preceding circumstances, we might just treat that as a certainty. That variable can then be removed from our chain of considerations, and we can then reason directly from congress' decision to the behavior of the SEC regulators.

The present problem, where the assessment of the probability of Y or X being the source of the latest missile launch is used to update propensities, can also be considered a multistage inference. When they come to update propensities, "Y only" responders may be treating the probability that Y launched the missile as a certainty. There is some evidence from the data to suggest they may be doing this. We found tentative evidence that "Y only" responders provided a higher mean estimate for the probability that Y launched the missile than "No change" responders (77.7% vs. 68.1%). It may be that their higher initial estimate makes them more willing to approximate to 100% that Y launched the missile. Indeed, Gettys et al. (1973) theorized that individuals would have a certainty threshold for doing this. The open text data are consistent with, but not specifically supportive of, this cognitive process, as individuals only seem to express their "belief" that Y launched the missile, and do not explicitly state the next step (how this leads to categorical propensity updating). The belief that the event cannot normatively be split, and "as if" reasoning could both theoretically constitute the "next cognitive step" following the first step in reasoning, and we will aim to determine which is the case in experiment 2. A final possibility which we consider unlikely but should be checked is that individuals have changed to become genuinely certain that Y launched the missile between the moment that they assessed this question and the moment they update the propensities.

Finally, this first experiment found "No change" responses in much larger numbers than initially anticipated, and the following experiment will seek to develop a greater understanding of the cognitive processes underlying this approach. Data from this first experiment indicate that uncertainty over the source of the missile seems to be a key factor. "No change" responders expressed lower certainty that Y was the source of the missile compared to "Y only" responders, and the most common code assigned in the qualitative analysis was "Don't know who launched the missile."

3. Experiment 2

In this experiment, we seek to further probe the reasoning of both the "Y only" and "No change" responses, the latter of which was not anticipated in such numbers in experiment

1. The open text data gave some indication of the cognitive processes underlying these responses; however, participants were not forthcoming about every aspect. We will, therefore, present these participants with a range of follow-up closed option statements (participants who give each response type will be funneled into a different set of questions) and request them to endorse that which most closely matches their reasoning. For the “No change” response, Anderson (2003) described the possibly relevant phenomenon of “decision avoidance” and laid out several cognitive and affective strategies which may underpin this. We aimed to determine if this response is due to a genuine belief that it is normative not to change the propensities based on uncertainty, or if this is due to a conservative strategy to play it safe, or finally due to avoiding regret about making the wrong decision (i.e., overupdating Y). We also aimed to check if this response was due to seeing the change as negligible (i.e., using no change as an approximation of negligible change), or as seeing the ME-propensities as fixed values that cannot change.

To supplement the closed responses, we will also provide “Y only” responders only with a new open text response opportunity to explain their reasoning, as in the first set of open text data, the final step in their reasoning (i.e., why they don’t give the “Both, Y more” response) was not explicit. To give the best chance of obtaining data on this point, we intend to present them with their response to the assessment of the evidence (i.e., demonstrate that they have provided a probabilistic assessment) and then their response to the posterior question (i.e., demonstrate that here they have provided a categorical response) and ask them explicitly to explain how they account for this apparent inconsistency. Finally, in order to investigate whether there is a threshold quality to the “Y only” response, as predicted by Gettys et al. (1973), these participants will be presented with a final follow-up question asking if they would have made the same response for a range of different levels of certainty that Y launched the missile this time.

3.1. Method

3.1.1. Participants

Participants for experiment 2 were recruited from Amazon MTurk and were required to have completed at least 50 tasks on the platform previously, have a 90% approval rating, and be based in the United States. Out of 256 total participants, 48.0% self-reported as female, 51.6% as male, and 0.4% as other. Mean age was 36.5 (SD = 11.8) with a minimum of 20 and a maximum of 71. Participants were asked their highest education level, and 33.6% reported high school, 49.2% bachelor’s degree, 11.3% master’s degree, 1.2% doctoral degree, and 4.7% other. Participants were also asked their profession and 42.2% reported “professional/managerial,” 33.2% “labor/service,” 6.6% “student,” 10.2% “unemployed,” and 7.8% “other.”

3.1.2. Design

Experiment 2 used the same scenario as experiment 1. Participants all had the same experience during the main part of the experiment; however, follow-up questions were presented to participants depending on their response to the key “propensity” question. Notably, those who increased their ME-propensity estimate for only Y, leaving X unchanged or reduced ("Y only"), were shown one set of follow-up questions, while those who made no change to either were shown another set of follow-up questions.

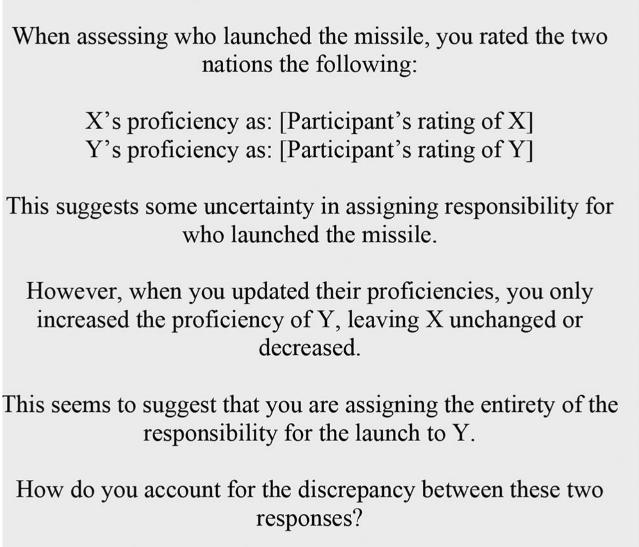
3.1.3. Procedure and materials

All participants were presented with the background information. They were then asked to provide an estimate of each nation’s ME-propensity in percentage form on a sliding scale (one for each nation). All participants were then presented with information regarding the missile explosion of an unknown source.

Following this, all participants were asked to provide an estimate as a probability that Y or X launched the latest missile (the same as the “Probabilistic separate” condition from experiment 1), and on a separate page they were then asked to indicate whether they would like to reduce/make no change/increase their estimate of the ME-propensity of Y and X. This response was collected using a 21-point sliding scale with three labels: “Decrease a lot” at the far left, “Make no change” in the middle, and “Increase a lot” at the far right. We increased the number of points and reduced the number of labels compared to experiment 1 as we wanted to ensure that the number of “No change” responses was not due to participants seeing the change as negligible. On the new slider, one “notch” on the scale is worth less than in experiment 1. On a separate slider, participants were also asked for their confidence in these two responses.

After providing their posterior ME-propensity updates, a range of follow-up questions were shown to participants depending on which of the two major response types they provided (“Y only” or “No change”). Principally, these involved the participants being shown a series of statements and being asked to endorse that which most closely resembled their thought process while providing the posterior response. Statements for “Y only” were developed based on previous theoretical work, such as Gettys et al. (1973), which provided inspiration for the “As if” statement. We also wished to pit this against a statement (“Cannot split”) testing whether it was the ultimate binary nature of the event which was causing participants to respond categorically. As a check that participants had not become certain since their response to the question about the single event, we also included a statement reflecting this (“Certain”). For the “No change” responder statements, Anderson’s (2003) work on indecision was used to inspire the “Safety” and “Regret” statements and these were pitted against a simple check of whether it is the uncertainty in the problem that causes this response (“Uncertainty-normative”), as well as checks that the participants did not see the propensity as either fixed, or the increase as negligible. All statements can be seen in the results section in Table 5. In some cases, choices led participants to another page with an additional set of statements—these can be seen nested within their parents within that table. Additionally, those providing the “Y only” response were randomly assigned in a 2:1 ratio to either see the closed choices in Table 5 (51 individuals) or to see an open text box (33 individuals) preceded by the instructions in Fig. 6.

Furthermore, all “Y only” responders were shown an additional follow-up question which requested them to indicate if they would still have increased their ME-propensity estimate of Y only, leaving X unchanged, if they had assessed the probability of who launched the missile differently. They answered this question for each value from 95% chance Y launched



When assessing who launched the missile, you rated the two nations the following:

X's proficiency as: [Participant's rating of X]
Y's proficiency as: [Participant's rating of Y]

This suggests some uncertainty in assigning responsibility for who launched the missile.

However, when you updated their proficiencies, you only increased the proficiency of Y, leaving X unchanged or decreased.

This seems to suggest that you are assigning the entirety of the responsibility for the launch to Y.

How do you account for the discrepancy between these two responses?

Fig. 6. Open text instructions provided to "Y only" responders in experiment 2.

the missile to 50% and were asked to choose between "Yes," "Not sure," and "No" at each level. This was included to check if there was a threshold effect to this response type as predicted by Gettys et al. (1973).

3.2. Results

3.2.1. Posterior responses

As can be seen in Fig. 7, out of the total sample of 256, 96 (37.6%) made no change to either X or Y and 84 (32.9%) provided the "Y only" response (increasing Y, leaving X unchanged or reduced). Other responses included increasing both, but Y more, given by 27 participants (10.5%) and increasing both equally, given by 13 participants (5.1%).

3.2.2. Prior, evidence, and confidence estimates

Participants were asked to provide initial estimates of X/Y's ME-propensity, before the latest explosion. The mean ME-propensity estimate for X was 24.0 (SD = 17.0) and for Y, the mean ME-propensity estimate was 68.9 (SD = 12.5).

Participant responses to the question asking the probability that X and Y launched the latest missile and their confidence in their posterior ME-propensity estimates can be seen in Table 4. We ran four linear regressions, each with "Y only" versus "No change" coded as a categorical predictor, and the variable from each row in the table as the criterion. From the top row, for "Launched Missile (X)," ($B = 3.9$, $F[1,178] = 4.1$, $p = .044$); for "Launched missile (Y)," ($B = 5.0$, $F[1,178] = 3.4$, $p = .068$); for "ME-propensity update Confidence (X)," ($B = 7.5$, $F[1,178] = 3.4$, $p = .068$) and finally, for "ME-propensity update confidence (Y)," ($B = 4.5$, $F[1,178] = 1.5$, $p = .225$).

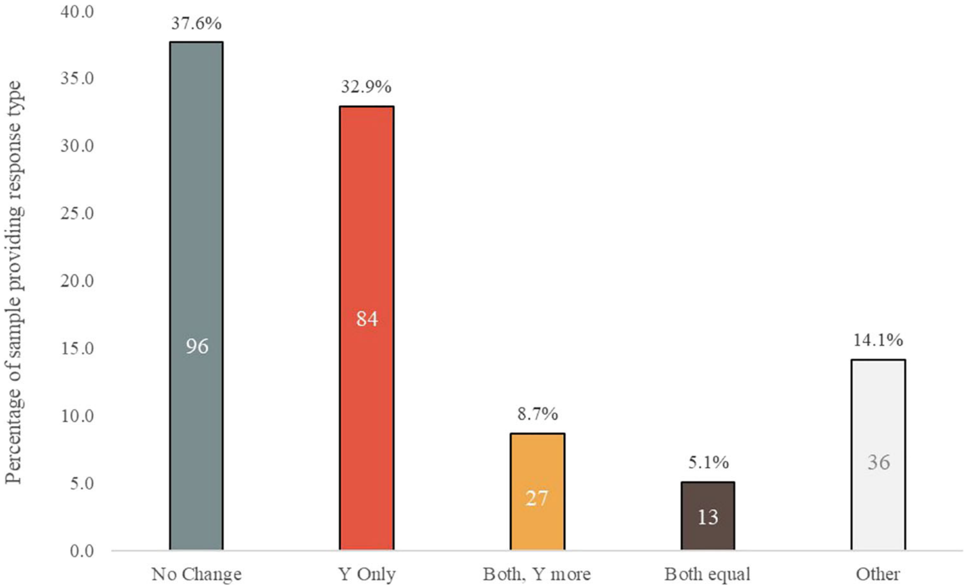


Fig. 7. The percentage of participants providing each response type in experiment 2.

Table 4
"Y only" and "No change" mean (SD) responses to question 1 and their confidence in their ME-propensity updates

	Y only (84)	No change (96)
Launched missile (X)	18.0% (11.9)	22.0% (13.9)
Launched missile (Y)	75.9% (19.4)	70.9% (16.9)
ME-propensity update confidence (X)	55.7% (27.6)	63.3% (27.4)
ME-propensity update confidence (Y)	71.6% (22.3)	67.1% (26.9)

3.2.3. *Closed choice responses*

Responses to a range of closed option statements by "Y only" and "No change" responders can be seen in Table 5. Out of those providing the "Y only" response, two-thirds (51) were assigned to see four closed options intended to reflect possible cognitive processes based on the theories provided so far (the remaining third were assigned to the open response format in Fig. 6). The number of individuals endorsing each statement can be seen as well as the percentage within the response's immediate parent category. All "No change" responders were assigned to see closed option statements and those who chose the "Uncertainty" statement in the first set of options were shown a further set of three options on a following page. These can be seen nested below that response. Responses are ordered from top to bottom at all levels according to their frequency.

Table 5
Statement endorsement frequencies and percentage of parent category for "Y only" and "No change" response types as well as for follow-up statements

	Freq	Parent%
Total sample	256	100
"No Change"	**96**	**37.6**
1. [Uncertainty] "You don't know who conducted this explosion"	64	66.7
1a. [Uncertainty-Normative] "The evidence states it's uncertain who launched the successful missile so you cannot change the proficiencies based on uncertainty"	52	81.3
1b. [Safety] "I am not sure it's correct to make no change to my estimates of X and Y's proficiencies but since I am unsure, it's the safest decision"	10	15.6
1c. [Regret] "I am afraid I might regret my decision if I make a change for one/both of the nations"	2	3.1
1d. Other	0	0
2. [Negligible] "A single extra observation makes a negligible change"	21	21.9
3. [Fixed] "An observation cannot change their actual proficiency"	11	11.5
4. Other	0	0
"Y only"	**84**	**32.9**
Closed options route	51	60.7
1. [Cannot Split] "Either Y OR X must have launched the missile, not both, so it would be incorrect to divide responsibility between them. I attributed the launch to the most likely source (Y)"	29	56.9
2. [As if] "I approximated that Y was entirely responsible for the launch in order to make the problem simpler, but know this is not strictly accurate"	16	31.4
3. [Certain] "I am now entirely certain Y launched the missile and X did not"	5	9.8
4. Other	1	2.0
Open text response route	33	39.3

Note. Brief indicators of the nature of the statement are given in bold in square brackets but were not shown to participants.

3.2.4. *"Y only" open text response*

Out of the 84 individuals who made the "Y only" response, 33 were randomly assigned to provide an open text explanation for their response after seeing Fig. 6, which attempted to make clear the inconsistency between their probabilistic assessment that Y launched the missile and their categorical updating of ME-propensities for X and Y. These were coded using the same five codes (plus "unclassified/other") and the same process outlined in experiment 1, and the final coding frequencies can be seen in Table 6.

Consistent with experiment 1, a majority of "Y only" participants were coded as explaining their approach by referencing the historical propensities and stating their (probabilistic) belief that Y launched the missile. For example, P25 wrote "Since Y was at 80%, I felt it was safe to assume that they launched the missile" and P26 wrote "Y has had a better success record so I feel confident that they were the one behind the launch." However, what is still missing here, despite our attempt to prompt it (Fig. 6), was an explanation of why they did not increase their estimates of both Y and X, but just Y more, in proportion to their certainty.

Table 6
The final agreed coding proportions for the five main codes, as well as "Unclassified/other" for "Y only" responders

	n	Historical propensities	Probably Y	Don't know who launched it	More data/ information needed	Nothing has changed	Unclassified/ other
"Y only"	33	70.0%	70.0%	3.0%	0.0%	0.0%	9.1%
Inter-rater *R*		81.8%	84.8%	97.0%	97.0%	100%	93.9%

Note. Initial inter-rater reliability for each code can also be seen in the bottom row.

3.2.5. Certainty thresholds for "Y only" responders

All 84 "Y only" responders were also shown a further follow-up question, requesting whether they would have made the same posterior response (increasing Y only) if their estimate of Y's probability of launching the missile had been different. Each participant chose between "Yes," "Not sure," and "No," and made this choice for a range of values (95, 90, 85, 80, 75, 70, 65, 60, 55, and 50). This was intended to determine if there was a certainty threshold aspect to this response, that is, if below a certain point, they would no longer consider it "safe to assume" Y was responsible.

We also considered it possible that those participants choosing "Cannot split" and those choosing "As if" may respond differently to this question. If you believe that the event cannot be split, you may be willing to make the "Y only" response at any value above 50%. However, since "As if" was intended to reflect a rounding process, these participants may only be willing to make the "Y only" response above a certain value. To test this, we first recoded these variables such that Yes = 1, "Not sure" = 0, and "No" = −1. This created 10 within-subjects variables, one for each value of Y (e.g., "95, 90 … 55, 50"). We entered these into a repeated measures ANOVA and also included a categorical between-subjects variable to indicate participants' statement choice ("Cannot split" vs. "As if"). We found a clear effect of "Value of Y" ($F[9,387] = 4.3, p = .00002$), no effect of statement choice ($F[1,43] = .07, p = .80$), and a small potential interaction ($F[9,387] = 1.9, p = .05$).

These effects can be seen in Fig. 8. "Cannot split" is represented by the dashed line, and "As if" by the dotted line. As can be seen, both follow a similar trajectory: although the small interaction effect might be picking up that "As if" starts lower but ends higher than "Cannot split," they do not seem to differ in the way theorized above. All "Y only" responders are, therefore, combined in the solid black line. Tentatively, this does appear to show relative stability with a proportion of around 0.7 saying "Yes" (they would make the "Y only" response) until roughly the 75% point after which linear decline is seen until "Yes" responses are below a proportion of 0.3 at the 50% point.

3.3. Discussion

In this second experiment, we have found consistent results with experiment 1, with very similar percentages of participants providing the "Y only" and "No change" responses (roughly one-third each). We have probed the reasoning of these two response types, which

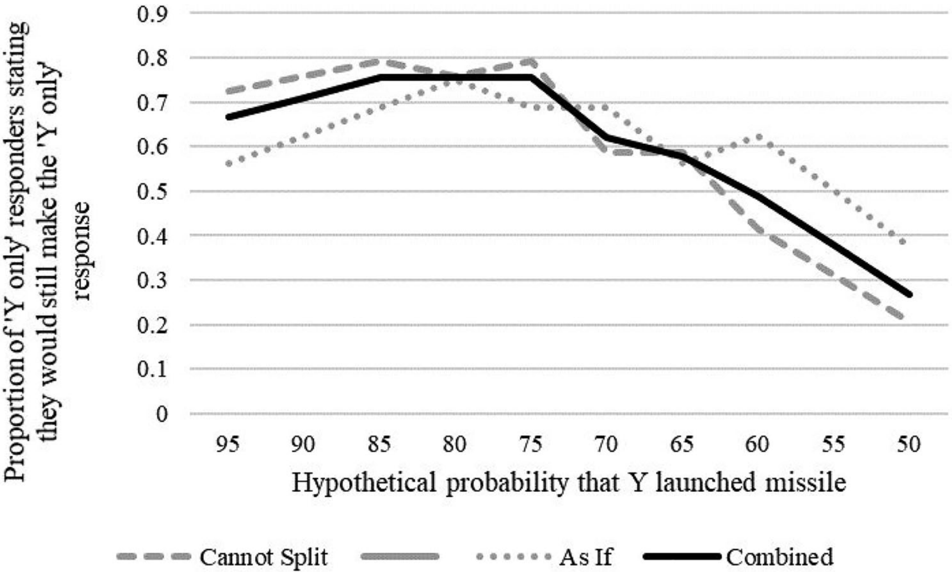


Fig. 8. The proportion (vertical axis) of "Y only" responders stating "Yes," that they would have made the same response if their assessment that Y launched the missile was different (values of the horizontal axis).

has provided some clarity but also opened up new questions. We attempt a synthesis of our current understanding of the cognitive processes underlying these two responses at this point.

3.3.1. Y only

Reinforcing the findings of experiment 1, the "Y only" response does not seem to be due to "black and white" thinking, in which participants would be certain that Y was responsible for the latest launch. Participants making this response again readily assign a probability to the chance that Y launched the latest missile, with a mean around 75%. Instead, it appears to be at the propensity updating phase that a probabilistic representation of the problem becomes categorical, or "digitized" (Gettys et al., 1973; Johnson et al., 2020). Through closed option responses, we explored two main reasons for this: "Cannot split" and "As if." "Cannot split" reflects the idea that because the missile launch is digital, that is, it must have been launched wholly by either X or Y, we cannot update ME-propensities in a graded way which appears to split responsibility between them. This runs in contrast to the Bayesian model, which splits the ME-propensity updates based on our certainty that each launched the missile. "As if" was designed to reflect a simple rounding tendency due to the difficulty of the problem, in line with Gettys et al. (1973). "Cannot split" received the greatest support from these participants (56.9%), but "As if" also received considerable support (31.4%). However, these two sets of participants responded very similarly to the "certainty threshold" question, with fewer participants stating they would make the "Y only" response as their certainty that Y launched the latest missile declines. This may suggest that these two clusters of participants are not entirely separate. It may be that the "cannot split" belief is a core driver of the "Y only" response (and perhaps in the scenarios studied by Gettys et al. also); however, participants recognize that this response involves an element of rounding ("As if")—participants may then be picking between these two statements based upon which aspect of their reasoning is more salient to them. Generally, it seems that the Y only response is very dependent upon the participants' certainty that Y is responsible. This fits with many "Y only" responders stating in their open text data that they felt "confident" enough that Y launched the missile to "assume" that they had launched it.

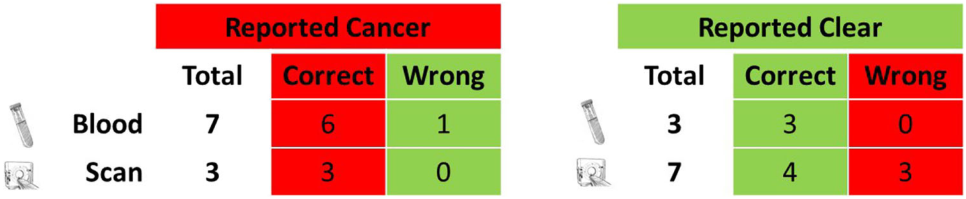
3.3.2. No change

"No change" responders were shown two stages of closed option questions designed to reflect a range of reasoning processes. In the first, they were asked to choose between "Uncertainty" ("You don't know who conducted this explosion"), "Negligible" ("A single extra observation makes a negligible change"), and "Fixed" ("An observation cannot change their actual proficiency"). While there was some support for "Negligible" (21.9%) and "Fixed" (11.5%), "Uncertainty" was the majority choice (66.7%). Participants who chose "Uncertainty" were shown a series of further options. The overwhelming choice among these was "Uncertainty-Normative" ("The evidence states it's uncertain who launched the successful missile so you cannot change the proficiencies based on uncertainty": 81.3%). Note that "proficiencies" is the term we used with participants for ME-propensity in the experiment.

3.3.3. Y only and No change

We have tentatively found that there may be a threshold aspect to "Y only' responders" approach to the problem: they seem to think it (1) not possible to split the ME-propensity update between both Y and X and (2) to think it likely enough that Y launched the missile to "assume" this at the ME-propensity-updating phase. We will call this the "categorical" approach to the propensity updating problem. However, it leads to the question of what these responders would do if they did not make this response, that is, if they were not "certain enough" Y was responsible to assume it was them. Would they make a graded response (which would seem inconsistent with the "cannot split" belief) or would they in fact jump all the way to the "No change" response?

Moving to the "No change" response, the closed option responses that we provided just used the term "uncertainty" which could refer to anything greater than 0 or less than 100. However, on reflection, this is relatively crude, and worth considering what these participants would do if they were more certain. Tentatively, "Y only" responders seem to be on average a bit more certain (77.7% in experiment 1, 75.9% in experiment 2) than "No change" responders (68.1% in experiment 1, 70.9% in experiment 2). What would "No change" responders do if they were more certain? It is hard to imagine that they would continue to make no change even if they were 99% certain. This leads to the possibility that "No change" responders are also operating on a threshold basis but are simply below their certainty threshold: they are either less certain than "Y only" responders, or they have a higher certainty threshold for making that response. In that case, it may be that these two "digital" responses to the problem are actually part of the same larger response. It may be that both hold the "cannot split" belief, that is, that if ME-propensities are to be updated, it must be only either Y or X, not both, and



	Reported Cancer			Reported Clear		
	Total	Correct	Wrong	Total	Correct	Wrong
Blood	7	6	1	3	3	0
Scan	3	3	0	7	4	3

Fig. 9. The exact stimulus presented to participants, displaying the accuracy figures for both tests for both types of report.

the only difference between them is that "Y only" responders are above their certainty threshold for ascribing this to Y, while "No change" responders are below their threshold. This possibility will be explored in experiment 3.

4. Experiment 3

In experiment 3, as well as addressing and extending the exploration of the two modal responses in experiment 2, we introduce a new, novel problem. In order to test whether the findings of experiments 1 and 2 generalize, we first change the causal structure (from a common effect structure to a common cause), and we move from a scenario involving agents and responsibility and potentially blame to a mechanistic scenario involving neither. This could make a difference as there is some evidence that binary thinking differs between moral and nonmoral domains (Johnson, Murphy, Rodrigues, & Keil, 2019) and in terms of the presence of "threat" (Zhu & Murphy, 2013). The problem is set during a clinical trial for two early warning tests for cancer: one using a blood test, and one using a scan test. So far, 10 patients have been through the full trial, first being tested with both and then monitored for 20 years to determine if the test result was correct, yielding the results in Fig. 9, which was presented to participants.

There are several patterns worth highlighting here. First is that in this group of 10 patients, six actually had cancer, and four were actually clear. Second, when the blood test does report cancer, it has been correct six times out of seven (six true positives, one false positive). Finally, when the scan test reports clear, it has been correct only four times out of seven (four true negatives, three false negatives). These particular numbers are important because in the next phase of the problem, participants are told (1) that a new high-risk patient has come forward (and that high-risk patients have a 50% chance of developing cancer over the next 20 years) and (2) both tests are run, with the blood test reporting cancer, and the scan test reporting clear. In this conflict scenario, where the two tests conflict with each other, participants' primary aim is to update the accuracy estimates for each test for those particular reports. To justify this, they were told that, since it takes such a long time to get confirmation (potentially 20 years), we need to keep the most precise estimate of these tests' accuracy updated at all times. This amounts to the "propensity to be correct when reporting cancer" for the blood test and the

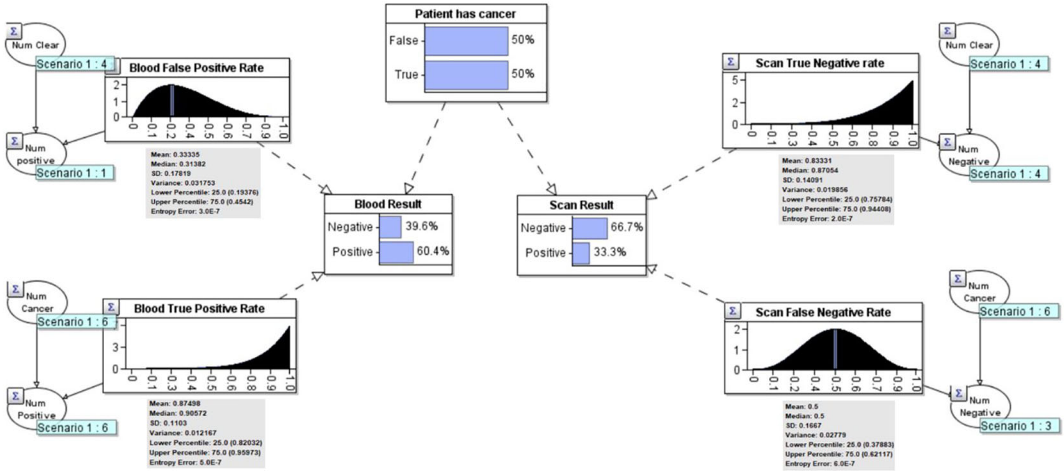


Fig. 10. A Bayesian network depicting the medical scenario before any observations.

"propensity to be correct when reporting clear" for the scan test. For the sake of fluency, we will just call this "accuracy" for those particular report types.

4.1. Bayesian model

These data can be turned into a Bayesian network model, similar to the missiles scenario. This can be seen in Fig. 10. The core of the model, and its common cause structure, can be seen in the three central nodes: "Patient has cancer" is a cause of both the "Blood Result" and the "Scan Result." However, the outcomes of these tests are also affected by their error rates which are depicted on the periphery. The blood test rates are on the left and the scan test rates are on the right. For ease of reading, we present the true-positive and false-positive rates for the blood test (because in this scenario that test reports a positive [i.e., cancer]), while for the scan test, we display the true-negative and false-negative rates (because that test reports a negative [i.e., clear]). The true-negative is the complement of the false-positive rate (i.e., if you have a false-positive rate of 5%, you have a true-negative rate of 95%) and the false-negative rate is the complement of the true-positive rate. This, therefore, makes no mathematical difference to the model but is instead simply a display choice which makes it easier for a human reader to understand the impact for both tests for the particular combination of observations we will make in this scenario.

Starting at the bottom left, we can see that the blood test's true-positive rate is based on 6/6 (87.5%): it did not miss-label any of the six patients who actually had cancer. Note that, just as in the missiles scenario model, the estimates here are based on a uniform prior and, therefore, produce estimates in line with the Laplacian rule of succession. This is a good example of a situation where this is necessary as without this adjustment the model is unable to handle frequencies like 6/6, which has no variance on a classical estimate. Moving to the top left, the blood test once mistakenly labeled a "clear" patient as having cancer, producing a false-

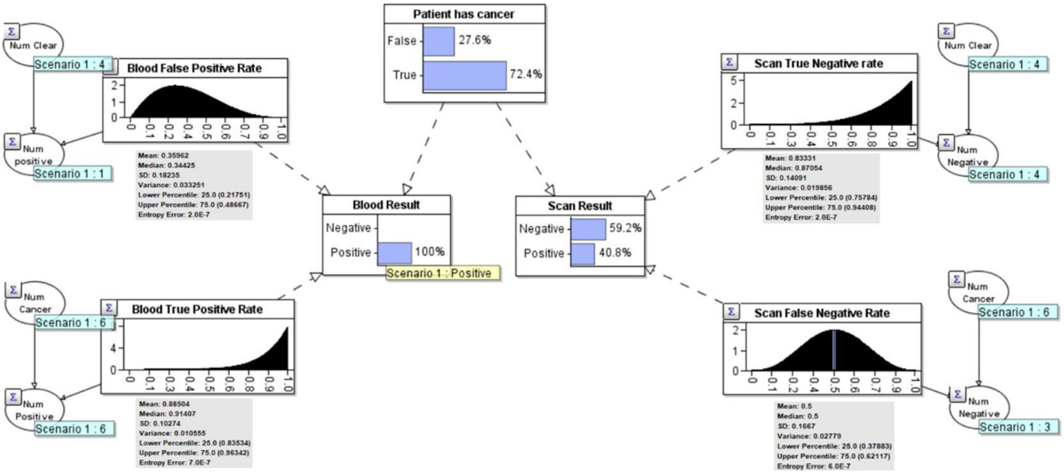


Fig. 11. A Bayesian network depicting the medical scenario after observing that the blood test reports positive.

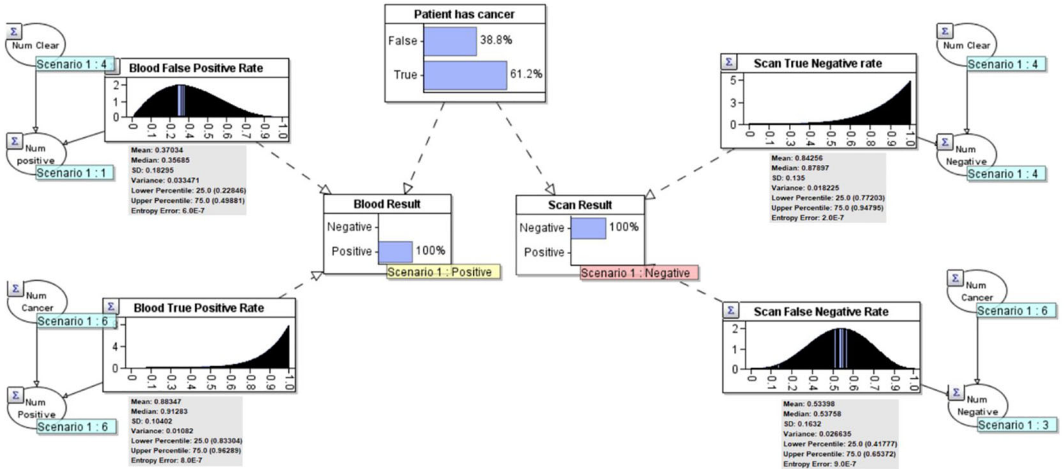


Fig. 12. A Bayesian network depicting the medical scenario after both the blood test reports positive, and the scan test reports negative.

positive rate of 1/4 (33.3%). Moving to the top right, the scan test correctly labeled all four clear patients as clear, producing a true-negative rate of 4/4 (83.3%). However, it did mislabel three of the patients with cancer as clear, producing a false-negative rate of 3/6 (50.0%).

We can make observations on both "Blood Result" and "Scan Result" and see what happens to the model's estimates of their accuracy for these particular reports. If we first observe that we just have the "Blood Result" test back (Fig. 11), and it says positive (the patient has cancer), we can see that both the true-positive and false-positive rates go up. This is because we know that one of these has occurred, but we do not know which.

We can then observe the negative "Scan Result" (Fig. 12). This is the scenario that the participants were asked to reason about: where the two tests are in conflict, but we do not

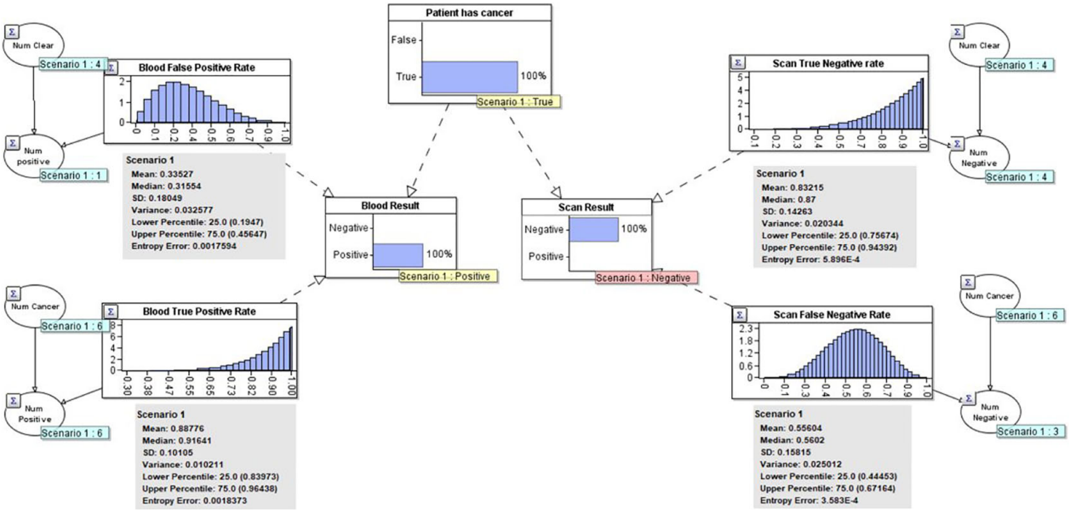


Fig. 13. A Bayesian network depicting the medical scenario after the blood test reports positive, the scan test reports negative, and we received confirmation that the patient really did have cancer.

know which is correct. First, looking at the blood test, we can see that the true-positive rate has almost come back down to its original level (∼1% higher than baseline), before any observations, while the false-positive rate has increased further (∼4% higher than baseline), reflecting the greater chance that this is now a false positive. We can also see that the false-negative rate of the scan result is considerably higher than baseline (∼4%), while the true-negative rate is only a little higher (∼1%). Therefore, regardless of the exact degree, under this conflict situation, both tests are seen as less accurate for these particular reports than before. In both cases, the ratio between their chance of making a false report versus a true report for these particular reports (positive for the blood test; negative for the scan test) is now higher.

It should be noted that participants were not asked to wrestle with these error rates, as this was deemed likely to produce confusion. Instead, having initially spent time familiarizing themselves with the numbers in Fig. 9, they were just asked to indicate whether they thought each of the two tests was now more accurate, less accurate, or the same, for these particular reports, just like experiments 1 and 2.

4.2. *Categorical updating*

In the first two experiments, we were able to definitively link categorical updating to a single pattern of responses. As shown in Fig. 4, if we were certain Y launched the missile, the normative response would be to increase the ME-propensity estimate for Y and leave X either unchanged or reduced (the "Y only" response). We can also determine what the normative response would be under certainty for the present medical scenario. This can be seen in Fig. 13. If we were certain that the blood test was correct (which we achieve by "observing" that the patient actually has cancer), we would increase our estimate of its accuracy for positive reports (as we would know it was correct) and decrease our estimate of the accuracy of the scan test for negative reports (as we would know it was wrong). We will call this the "Increase-Reduce" (or "Inc-Red") response.

However, unlike in experiments 1 and 2, this response pattern is not so exclusively associated only with the response under certainty. While the particular set of frequencies that we used in the medical problem leads to reduced accuracy for both tests for these types of reports under the "uncertain conflict" situation (Fig. 12), the pattern of updates depends on the particular frequencies used. This is because the same dynamic exists here as with the missiles model, where the closer that the mean of any of the four "rates" is to 100%, the smaller the increase once the two reports are observed. In the example presented in Fig. 12, both the false-positive rate of the blood test and the false-negative rate of the scan test increase a lot more than the true-positive and true-negative rates (i.e., both tests become less accurate), in part because the mean of those rates is much further from 100%. However, with other frequencies, situations can be found where either the accuracy of both increase (i.e., the "true" rates increase more than the "false" rates), the blood test accuracy increases more than the scan test, or even that the scan test accuracy increases more than the blood test (see the online repository for examples and discussion).

For this reason, a common cause scenario cannot as definitively determine, just based on the pattern of propensity updates, whether participants are using the categorical approach or not. While the frequencies we used do lead to the accuracy of both tests decreasing, we do not expect participants to be sensitive to the subtle mathematical dynamics that produces this result (i.e., decreasing updates as the mean estimate approaches 100%) and so we must consider the "Increase-Reduce" response a reasonable response to the problem, even with these frequencies. In this third experiment, we, therefore, include a direct question asking all participants if they assumed that one of the two tests was correct when updating accuracy estimates.

Furthermore, in order to develop the theory of the relationship between "Categorical" and "No change" responses, both sets of participants were asked, first, their certainty that the blood test was correct on this occasion, and second, their certainty thresholds for increasing their estimate of the blood test (in experiment 2, this was only elicited from "Y only" responders). No change responders were also asked a new set of closed option questions to probe their reasoning.

Finally, to avoid any issue of preconceptions around blood tests versus scan tests, the scenario was in fact run in two versions: one with the numbers shown in Fig. 9, and two, with the numbers completely reversed. In that version, the blood test still reported cancer, while the scan test reported clear, however, the numbers were completely flipped so that, for example, the scan test had the 6/6 true-negative rate (the exact figures for both versions can be seen in Table 7). For this reason, in the following, we will refer to the more historically accurate (HA) test, rather than specifically to the blood test.

Table 7
The provided figures and participant estimates for both report types, for both tests, and for both versions of the experiment

Version:	Blood better		Scan better	
	Provided	Mean (SD)	Provided	Mean (SD)
Blood-Cancer	6/7	76.3 (21.3)	4/7	56.8 (12.8)
Blood-Clear	3/3	85.8 (25.5)	3/3	84.8 (22.3)
Scan-Cancer	3/3	82.5 (26.9)	3/3	86.1 (22.0)
Scan-Clear	4/7	53.8 (18.1)	6/7	80.2 (14.8)

4.3. Method

4.3.1. Participants

Participants for experiment 3 were recruited from Prolific Academic. Out of 225 participants, 44.4% self-identified as female, and 55.6% self-identified as male with no one choosing the “other” option. Mean age was 28.8 (SD = 10.4) with a minimum of 18 and a maximum of 75.

4.3.2. Design

Similar to experiments 1 and 2, all participants were presented with the same scenario, but experienced different follow-up questions based on their responses.

4.4. Materials and procedure

Participants were presented with the basic scenario as described in the introduction and were presented with the numbers in Fig. 9. They were then asked to provide initial accuracy estimates for both tests, for both types of reports (four estimates total).

Following this, participants were told that there is a new high-risk patient (with a 50% chance of developing cancer), then that the blood test reported cancer, and then that the scan test reported clear. Finally, both reports were presented, and participants were reminded of the original statistics from Fig. 9 as well as their own initial estimates for these particular reports only.

Participants were then asked to indicate, on the same type of scales used in experiment 2, whether, in light of these reports, they see the blood test, for cancer reports (and separately, the scan test, for clear reports), as less accurate, more accurate, or still the same as their original estimate. Their original estimate was “piped in” next to “Still” so, for example, a participant may see “Still 75%” if they originally estimated 75%. Beneath this, participants were presented with a single open text box and asked to explain their responses.

Following this, all participants were asked to indicate their certainty “that the blood test [scan test, in the counterbalanced condition] is correct this time” as a sliding scale from 0% to 100%, as well as their confidence in that estimate, also from 0% to 100%. Participants were then asked, “When deciding how to update your accuracy estimates of the two tests, did you”:

Table 8
Participant responses to four questions divided by the two major response types: No change and Inc-Red

		No change	Inc-Red
	Total *N* (225)	73	75
Mean (SD)	**Certainty** more HA test is correct	74.7% (18.7)	76.3% (15.0)
	Confidence in above	66.0% (20.4)	72.8% (18.7)
	Certainty threshold for increasing accuracy estimate of more HA test	84.1% (12.4)	76.4% (12.1)
Proportion (self-reported)	Assumed more HA test correct	52.1%	81.3%
	Assumed neither/other	38.4%	12.0%
	Assumed less HA test correct	9.6%	6.7%

and chose from options "Assume the blood test was correct this time," "Assume the scan test was correct this time," and "Neither/Other."

Following this, in order to elicit minimum certainty thresholds for increasing their accuracy of the more HA test, "Inc-Red" and "No change" responders were asked to pick a minimum certainty from a set of multiple-choice options ranging from 50% to 100%. This was unlike experiment 2 where "Y only" responders were asked to indicate, for every 5% interval, whether they would have made the same response or not. This new format was designed to be both as intuitive as possible, but also to allow us to easily calculate a mean threshold for each response type.

4.5. *Results*

The mean initial estimates for the accuracy of each test, for each type of report, can be seen in Table 7. First, for the 3/3 estimates, we can see a consistent mean around 85, with considerable spread (SD = ~25). For the 6/7 figure, estimates are 76.3 and 80.2 and for the 4/7 figure, 53.8 and 56.8, so we can clearly see that in both versions of the experiment, participants clearly accepted that one test was more historically accurate (HA) for the particular types of reports seen.

Similar to experiments 1 and 2, there were two modal responses to the problem based on the way participants updated their estimates of the accuracy of the two tests. The first was "Inc-Red," where participants increased their estimate of the accuracy of the more-HA test for the particular report made (e.g., their accuracy for reporting cancer, for the blood test; clear for the scan test) while decreasing their estimate for the other test. This is tentatively equivalent to "Y only" in the first experiment, as it would be the correct response if you were 100% certain that the blood test was correct this time (as demonstrated in Fig. 13). The second was "No change" where participants made no alterations to their estimates for the accuracy of either test. No other response type was seen in more than 4% of participants. In Table 8, summary statistics for the answers to a range of questions are shown for each of these two responses, for comparison.

For the subsequent analyses, we created a simple binary predictor variable called "Response" with "No change," coded as 0, and "Inc-Red" coded as 1. In the top row of

Table 9
"No change" participant endorsements to a set of closed option statements designed to reflect different thought processes

"No change"	73	100.0
1. [Certainty] Until I know for certain whether the blood test is correct this time (i.e., whether the patient really has cancer) it is incorrect to make any change to my accuracy estimates.	44	60.3
2. [Fixed] I saw my estimate of the tests' accuracy as exactly equalling its true accuracy, which cannot change, whatever new information we get.	14	19.2
3. [Negligible] Although my accuracy would change a little, it is a negligible change from only one extra observation.	14	19.2
4. Other	1	1.4

the table, participants were asked, before they updated their accuracy estimates, for their certainty that the more HA test is correct about the current patient. As can be seen, certainty levels are quite similar (No change = 74.7%; Inc-Red = 76.3%). We ran a linear regression with Response as predictor and Certainty as criterion ($B = 1.6$, $F(1,146) = .334$, $p = .564$).

In the second row of the table, participants were asked for their confidence in the above certainty estimate. Some difference can be seen in confidence between these two responses, with Inc-Red (72.8%) showing higher confidence levels than No change (66.0%). We ran a linear regression with Response as predictor and Confidence as criterion ($B = 6.8$, $F(1,146) = 4.5$, $p = .035$).

In the third row of the table, participants were asked, after updating their accuracy estimates, what certainty they would require (their threshold) for increasing their estimate of the accuracy of the more-HA test. A larger difference can be seen here, with No change (84.1%) showing a higher threshold than Inc-Red (76.4%). We ran a linear regression with Response as predictor and Certainty threshold as criterion ($B = 7.7$, $F(1,146) = 14.2$, $p<.001$).

The final three rows of the table present answers to a single multiple-choice question. Participants were asked if, after updating their accuracy estimates, they had assumed either the blood test was correct, the scan test was correct, or neither/other. We can see that more "Inc-Red" responders self-reported assuming the more HA test was correct, while more "No change" responders self-reported assuming neither/other. We coded this into a single variable (Assume), assigning participants a $+1$ if they self-reported assuming the more HA-test was correct, 0 if neither/other, and -1 for the less HA test. To test for a linear effect, we ran a linear regression with Response as predictor and Assume as criterion ($B = 0.3$, $F(1,146) = 10.0$, $p = .002$).

Finally, "No change" participants were shown a set of option statements, similar to experiment 2 but with the first option flipping the "uncertainty" statement to asking about the need for certainty. As can be seen, the majority chose the option indicating that they needed certainty to update their estimates, but some support was again seen for "Fixed" and "Negligible" (Table 9).

Table 10
The final agreed coding proportions for the five main codes, as well as "Unclassified/other" for both "No change" and "Increase-Reduce" response types

	n	Historical propensities	More-HA test probably correct	Don't know which is correct	More data/ information needed	Nothing has changed	Unclassified/ other
No change	73	31.5%	2.7%	13.7%	15.1%	17.8%	37.0%
Inc-Red	75	65.3%	21.3%	1.3%	1.3%	0.0%	30.7%
Inter-rater *R*		84.5%	83.1%	93.9%	97.3%	87.8%	78.4%

Note. Initial inter-rater reliability for each code can also be seen in the bottom row.

4.6. Qualitative data

"Increase-Reduce" and "No change" responders' data were coded using the same process and the same five codes used in experiments 1 and 2. The results can be seen in Table 10.

As can be seen in Table 10, some strong patterns observed in responses to the "missiles" scenario are less convincing here and the proportion of "Unclassified/other" is slightly higher than in experiment 1. First, there are fewer individuals explicitly stating that they think one or the other test is correct. Instead, the majority of "Inc-Red" responders (65.3%) simply refer to the historical accuracies of the two tests when explaining their response. This is in conflict with the data from the question asking whether they had assumed that one of the two tests was correct on this occasion. It is possible, when comparing this to experiments 1 and 2, which saw much higher frequencies on this code for "Y only" responders, that the "responsibility attribution" nature of that problem may have led to more participants feeling the need to say whether they thought X or Y was responsible. When interpreting this sort of data, we must always consider Gricean (Grice, 1975) principles of conversation: participants only tend to write a few sentences and so do not say everything on their mind. They will choose what to include, and what not to include based on what they think is most important for the experimenter to know. Stating the outcome of the single event may have seemed more important to them in a common effect responsibility attribution situation than in this common cause medical situation. However, we do still see much higher numbers making this response among "Inc-Red" (21.3%) compared to "No change" (2.7%), and those who do make this response seem to state it in similar terms to experiments 1 and 2. For example, P11 said "They are in a higher risk group and the blood test has a better history of being accurate so it seems more likely that the blood test is right" and P76 said "The blood test is far more reliable so I think it being positive in a high-risk patient is almost certainly going to be correct."

Similarly, fewer "No change" responders received the "Don't know which is correct" code, and similar numbers were assigned to "More data/information needed" and "Nothing has changed." It should be noted that the situation here, where both tests had made their reports, is quite a different situation to the common effect situation where something concrete has happened (a missile has exploded). It is perhaps not surprising that here, with just the reports of the tests, that many participants would feel that "Nothing has changed." For example, P153 wrote “I still think they have the same accuracy as before because the blood and scan tests are independent of each other. Having reports from both tests available at the same time does not affect the individual accuracy of the tests” and similarly, P24 wrote “The results of each test are not dependent on the other, therefore I do not believe that the accuracy of the tests would change.”

4.7. Discussion

In experiments 1 and 2, we found tentative evidence that “Y only” responders may be more certain that Y was responsible than “No change” responders. In this experiment, certainty was much more clearly equal; however, we found that certainty thresholds were markedly different. In reality, a difference in either certainty or threshold could manifest in an experiment like this, as what we believe differentiates these two response types is the relationship between these two things (i.e., categorical responders are at or above their threshold, and no change responders are below).

It is difficult to be fully confident in the equivalence of the certainty threshold questions asked to “Inc-Red” and “No change” responders. This is because “Inc-Red” responders were being asked about a response they had made, while “No change” responders were being asked about one they had not. This should make this result tentative, and future work should examine this same question with a different approach to confirm. For example, a within-subjects design manipulating the frequencies in the problem may be valuable to see if individual participants actually flip between these two response types without ever giving a graded response.

5. General discussion

In this paper, we have introduced a new type of problem to the judgment and decision-making literature: the updating of propensity estimates based on uncertain instances. This type of problem asks participants to update their estimates of figures which have typically been assumed to be fixed in previous work (e.g., Bar-Hillel, 1980). The process of monitoring and updating propensities is a continual daily process for human beings and so worthy of further study. In this preliminary set of experiments, we have found, across two different scenarios with different causal structures, that many participants update their propensities in a binary manner, either updating as if they have complete certainty about the particular instance (the “categorical” response [“Y only” in experiments 1 and 2, “Inc-Red” in experiment 3]), or as if they have no information at all (the “No change” response). Particularly in the common effect “responsibility attribution” situation (experiments 1 and 2), the categorical response seems to have obvious parallels to real-life situations, for example where someone is assumed to be responsible for some negative event and is then seen as even more likely to produce such events (e.g., to misbehave) in future. It is important to note that while we have not found “black and white” thinking in this paper (perceived certainty that one party is responsible), this would undoubtedly occur in real-world situations with more emotional valence and would be another route to the same propensity updating pattern. This paper has, therefore, highlighted a circular cognitive mechanism by which beliefs about, for example, other agents can take on a positive-feedback dynamic when faced with uncertain new instances via the assumption that an agent with a history of causing such events is responsible on this occasion and then updating propensity estimates based on that assumption.

Across three experiments, we have explored participant responses when asked to update propensity estimates based on a new uncertain instance. In the first two experiments, participants updated their estimate of the propensity for the missiles of two nations to explode (ME-propensity) after being launched. This used a common effect structure, and participants were asked to update their estimates for both nations, both of which could have been the source of the launch. In the third experiment, participants were asked to update the propensity for two early warning cancer tests to be accurate when providing particular reports. This used a common cause structure, where both tests produced conflicting reports. In both experiments, one of the two propensities to be updated had a higher prior: in experiments 1 and 2, Y had a higher frequency of explosions in the past; in experiment 3, one test had been more accurate in the past. In both scenarios, we find a substantial number of participants giving what we call the "categorical" approach. This manifests differently in the two scenarios: in the missiles scenario, participants approaching the problem categorically update in a way which would be normative if they were 100% certain that Y is responsible (increase Y, leave X unchanged or reduced ["Y only"]). In the medical scenario, the categorical approach leads to updating in a way which would be normative if they were 100% certain that one of the tests is correct (increase the more HA-test, reduce the other ["Increase-Reduce"]). While the "Increase-Reduce" response in experiment 3 cannot be so exclusively linked to categorical updating, a large majority (81.3%) of participants making this response self-reported assuming that the more-HA test was correct when updating accuracy. Furthermore, the percentage of participants making this response was very similar (around one-third) to that in experiments 1 and 2, making us confident that at least a substantial proportion of "Increase-Reduce" responders are indeed reasoning categorically when updating propensities. The second major response type observed is to make no change at all to either propensity, as if the participant has no information. Unlike the categorical approach, this manifests in the same way in both scenarios. These two responses ("Categorical" and "No change"), therefore, represent two sides of a binary approach to the problem: as if we have complete certainty, or no information at all.

Over both experiments, we have used a range of methods to attempt to probe the reasoning of both of these responses and we attempt the following synthesis. Both of these situations reflect uncertainty about an event which must, in reality, be one way or the other (either X or Y launched the missile, not both; the patient either does or does not have cancer). The graded/probabilistic nature of both problems comes from our epistemic uncertainty about that event, not from the event itself being a graded phenomenon. In experiment 2, when asked to endorse the statement which most represented their thinking, 56.9% of "Categorical" ("Y only") responders chose the statement which reflected this: "Either Y OR X must have launched the missile, not both, so it would be incorrect to divide responsibility between them. I attributed the launch to the most likely source (Y)."

Across experiments 2 and 3, we probed "No change" reasoning through closed option questions. In experiment 2, a majority endorsed a statement that it was wrong to update estimates under uncertainty, and in experiment 3, a majority endorsed a statement that it was wrong to update until certainty was achieved. Complete certainty (i.e., 100%) may not be required, however. In experiment 3, we asked "No change" responders what certainty level they would need in order to increase their estimate of the more HA-test, and the mean was around 85%. Indeed, while "No change" responders' certainty that the more HA-test was correct this time was on average below their threshold (~74 vs. ~85), "Categorical" responders were on average at their threshold (~74 vs. ~74). We, therefore, suggest that both response types in fact hold a similar representation of the problem. We propose that both see these two responses as the only two possible responses to the problem (no graded approach in between is permitted), and that they are picking between the two based upon whether they feel "certain enough" to make the categorical response (i.e., whether their certainty is above their threshold): if they do not, they take the more conservative "No change" approach.

This conclusion is potentially concerning. On several occasions throughout this paper, we have highlighted the issue with the categorical approach to propensity updating. Based purely on historical propensities, it first assumes that the more probable interpretation of the single event is correct, and then updates those very same propensities based upon that assumption. Our synthesis suggests that, rather just one-third of our sample, in fact two-thirds of our sample would respond in this way if the frequencies were more convincing (e.g., if Y had 5/6 rather than 4/6 successful explosions) and that the only reason we do not see more categorical responses is because the certainty threshold of many participants has not been met. Unfortunately, our attempts in experiment 1 to reduce this type of response, by encouraging participants to think probabilistically about the single event were not successful. This may be because participants hold the "cannot split" belief which does not permit a probabilistic response and future interventions may be useful in attempting to target this belief.

This paper produces a bridge between literature demonstrating the human tendency to treat probabilities categorically or "digitally" (e.g., Gettys et al., 1973; Johnson et al., 2020) and the confirmation bias/belief polarization literature (Cook & Lewandowsky, 2016; Jern et al., 2014; Lord et al., 1979; Nickerson, 1998; Plous, 1991). Authors such as Jern et al. (2014) have shown that the typical confirmation bias process of (1) interpreting uncertain evidence based on priors and (2) updating those same priors based on that interpretation is in fact acceptable within a Bayesian framework. Our models show the same: for example, in the missiles scenario, with a completely uncertain new instance (no indicating evidence of the source other than historical propensities), the ME-propensity of the candidate source with the higher historical ME-propensity (Y) is updated more, all else being equal. However, our "categorical" responders take this one step further. These participants (1) interpret uncertain evidence based on their priors, and (2a) convert that interpretation into a certainty when (2b) updating their estimates of that very same prior. This approach is not acceptable within a Bayesian framework, which produces graded/probabilistic updates. Many studies of confirmation bias/belief polarization have not made this distinction between graded and categorical updating and so this result may be important in this debate about the "rationality" of confirmation bias/belief polarization.

Acknowledgments

Funding was partly provided by the ERC project ERC-2013-AdG339182-BAYES_KNOWLEDGE and the Leverhulme Trust project RPG-2016-118 CAUSAL-DYNAMICS.

Open Research Badges

This article has earned Open Data and Open Materials badges. Data and materials are available at https://osf.io/etsav/.

References

Anderson, C. J. (2003). The psychology of doing nothing: Forms of decision avoidance result from reason and emotion. *Psychological Bulletin*, *129*(1), 139.

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica*, *44*(3052), 211–233.

Birnbaum, M. (1983). Base rates in Bayesian inference: Signal detection analysis of the cab problem. *American Journal of Psychology*, *96*(1), 85–94.

Casscells, W., Schoenberger, A., & Graboys, T. B. (1978). Interpretation by physicians of clinical laboratory results. *New England Journal of Medicine*, *299*(18), 999–1001.

Chen, S. Y., Ross, B. H., & Murphy, G. L. (2014). Decision making under uncertain categorization. *Frontiers in Psychology*, *5*, 911.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*(2), 367–405.

Cook, J., & Lewandowsky, S. (2016). Rational irrationality: Modeling climate change belief polarization using Bayesian networks. *Topics in Cognitive Science*, *8*(1), 160–179.

Dewitt, S. H., Fenton, N. E., Liefgreen, A., & Lagnado, D. A. (2020). Propensities and second order uncertainty: A modified taxi cab problem. *Frontiers in Psychology*, *11*, 503233.

Dyson, S. B. (2009). Cognitive style and foreign policy: Margaret Thatcher's black-and-white thinking. *International Political Science Review*, *30*(1), 33–48.

Fenton, N., & Neil, M. (2018). *Risk assessment and decision analysis with Bayesian networks* (2nd edition). Chapman and Hall/CRC Press.

Frenkel-Brunswik, E. (1949). Intolerance of ambiguity as an emotional and perceptual personality variable. *Journal of Personality*, *18*(1), 108–143.

Fryer, R. G., Harms, P., & Jackson, M. O. (2015). Updating beliefs when evidence is open to interpretation: Implications for bias and polarization. *Journal of the European Economic Association*, *17*(5), 1470–1501.

Gallistel, C. R., Krishan, M., Liu, Y., Miller, R., & Latham, P. E. (2014). The perception of probability. *Psychological Review*, *121*(1), 96–123.

Gettys, C. F., Kelly, C., & Peterson, C. R. (1973). The best guess hypothesis in multistage inference. *Organizational Behavior and Human Performance*, *10*(3), 364–373.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102*(4), 684–704.

Glad, B. (1983). Black-and-white thinking: Ronald Reagan's approach to foreign policy. *Political Psychology*, *4*(1), 33–76.

Grice, H. P. (1975). Logic and conversation. In Syntax and Semantics: *Speech acts, Vol 3. New York: Academic Press* 41–58.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*(4), 334–384.

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

Johnson, S. G. B., Merchant, T., & Keil, F. C. (2020). Belief digitization: Do we treat uncertainty as probabilities or as bits? *Journal of Experimental Psychology: General*, *149*, 1417–1434.

Johnson, S. G. B., Murphy, G. L., Rodrigues, M., & Keil, F. C. (2019). Predictions from uncertain moral character. In A. K. Goel, C. M. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st Annual Conference of the Cognitive Science Society* (pp. 506–512). Austin, TX: Cognitive Science Society.

Kahneman, D., & Varey, C. A. (1990). Propensities and counterfactuals: The loser that almost won. *Journal of Personality and Social Psychology*, *59*(6), 1101.

Keren, G., & Teigen, K. H. (2001). The probability-outcome correspondence principle: A dispositional view of the interpretation of probability statements. *Memory & Cognition*, *29*(7), 1010–1021.

Laplace, P.-S. (1814). *Essai philosophique sur les probabilités*. Paris: Courcier.

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*, 2098–2109.

Murphy, G. L., & Ross, B. H. (1994). Predictions from uncertain categorizations. *Cognitive Psychology*, *27*(2), 148–193.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220.

Plous, S. (1991). Biases in the assimilation of technological breakdowns: Do accidents make us safer? *Journal of Applied Social Psychology*, *21*(13), 1058–1082.

Popper, K. (1959). The propensity interpretation of probability. *British Journal for the Philosophy of Science*, *10*(37), 25–42.

Rabin, M., & Shrag, J. L. (1999). First impressions matter: A model of confirmatory bias. *Quarterly Journal of Economics*, *114*(1), 37–82.

Sanborn, A. N., Noguchi, T., Tripp, J., & Stewart, N. (2020). A dilution effect without dilution: When missing evidence, not non-diagnostic evidence, is judged inaccurately. *Cognition*, *196*, 104110.

Tesic, M., Liefgreen, A., & Lagnado, D. (2020). The propensity interpretation of probability and diagnostic split in explaining away. *Cognitive Psychology*, *121*, 101293.

Welsh, M. B., & Navarro, D. J. (2012). Seeing is believing: Priors, trust, and base rate neglect. *Organizational Behavior and Human Decision Processes*, *119*(1), 1–14.

Yu, A. J., Dayan, P., & Cohen, J. D. (2009). Dynamics of attentional selection under conflict: Toward a rational Bayesian account. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(3), 700–717.

Zhu, J., & Murphy, G. L. (2013). Influence of emotionally charged information on category-based induction. *PLoS One*, *8*(1), e54286.

Zylberberg, A., Wolpert, D. M., & Shadlen, M. N. (2018). Counterfactual reasoning underlies the learning of priors in decision making. *Neuron*, *99*(5), 1083–1097.