The zero-sum fallacy in evidence evaluation

Toby D. Pilditch[1*], Norman Fenton[2] and David Lagnado[1]

University College London[1] & Queen Mary University of London[2]

*Corresponding Author: Correspondence should be addressed to Toby D. Pilditch, 26 Bedford Way, London, WC1H 0AP, UK. Electronic mail can be sent to t.pilditch@ucl.ac.uk.

Abstract

There are many instances, both in professional domains such as law, forensics, and medicine, and in everyday life, where an effect (e.g. a piece of evidence or event) has multiple possible causes. In three experiments we demonstrate that individuals erroneously assume that evidence which is equally predicted by two competing hypotheses offers no support for either hypothesis. However, this assumption only holds in cases where competing causes are mutually exclusive and exhaustive (i.e. exactly one cause is true). We argue this reasoning error is due to a zero-sum perspective on evidence, wherein people assume that evidence which supports one causal hypothesis must disconfirm its competitor. Thus, evidence cannot give positive support to both competitors. Across three experiments (*N* = 49; *N* = 193; *N* = 201) we demonstrate this error is robust to intervention and generalizes across several different contexts. We also rule out several alternative explanations of the bias.

Keywords: zero-sum, intuitive judgment, cognitive bias, evidential reasoning, probabilistic reasoning

Introduction

In 2001 Barry George was convicted of the shooting of Jill Dando, a TV celebrity, outside her flat in broad daylight. The main evidence against him was a single particle of firearm discharge residue (FDR) found in his coat pocket. In 2007 the Appeal Court concluded that the FDR evidence was not ‘probative’ in favour of guilt, because, contrary to what had been suggested in the original trial, it was equally likely to have arisen due to poor police procedures (such as the coat being exposed to FDR during police handling) as from him having fired the gun that killed Dando. Hence, his conviction was quashed and a re-trial ordered, in which Barry George was set free.

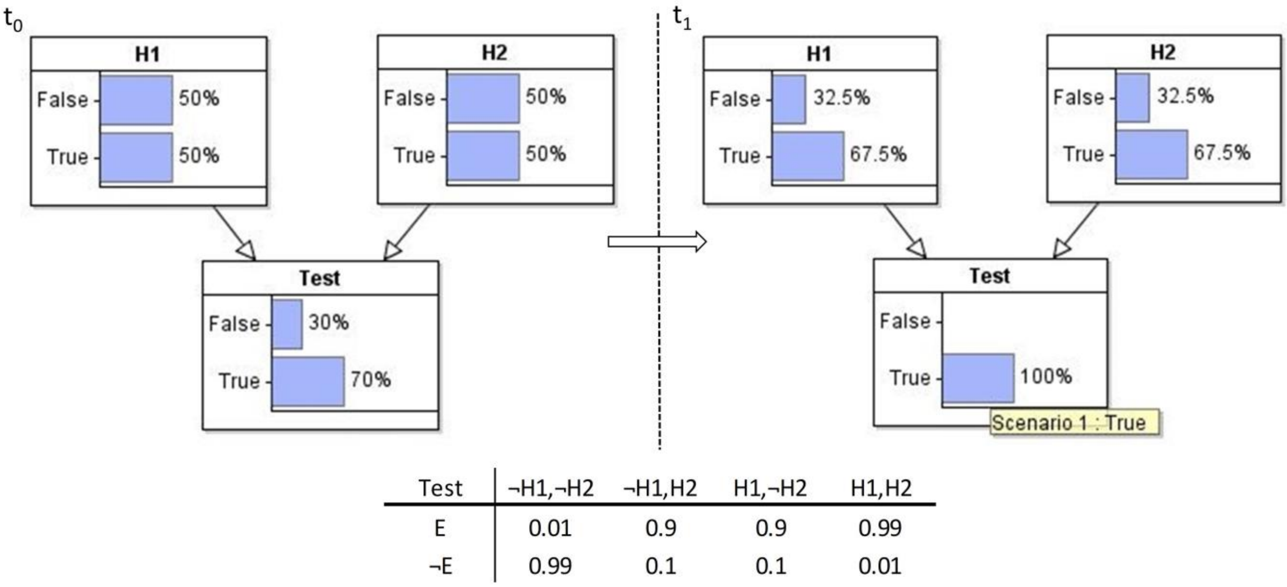
How valid was the court’s argument that the FDR was non-probative? Fenton et al (2014) show that the main argument presented in the appeal judgment may have been flawed: the argument assumed that if a piece of evidence (the FDR in the coat pocket) is equally probable under two alternative hypotheses (Barry George fired gun vs poor police handling of evidence) then it cannot support either of these hypotheses. But this assumption only holds if the two alternative hypotheses are mutually exclusive and exhaustive (i.e. exactly *one of these two* hypotheses is true). In the Barry George case this is clearly not met: it is possible that he fired the gun *and* there was poor police handling of the evidence; and also that neither were true (e.g., the FDR particle came from elsewhere). Therefore, rather than being neutral, the FDR evidence may have been probative against Barry George (albeit weakly). The FDR evidence does not discriminate ‘Barry George fired the gun’ versus ‘poor police handling of evidence’, but it does discriminate ‘Barry George fired the gun’ from ‘Barry George *did not* fire the gun’: it is the latter hypothesis pair that was the target in this criminal investigation.

This error was committed in the highly charged context of a criminal appeal, and involving legal and forensic experts. But it identifies a reasoning error that is potentially very pervasive, as it goes to the heart of standard methods for evaluating evidence in terms of likelihood ratios, and also arises informally in many contexts where evidence is evaluated. In this paper we demonstrate that the reasoning error is prevalent in everyday lay judgments about the value of evidence, and that it persists despite attempts to alleviate the bias through clarifying instructions. Furthermore, we show that people are perfectly capable of assessing the value of negative evidence, using it to "rule out" hypotheses accordingly. Finally, we propose a simple psychological mechanism that underpins our findings, based on the notion that explanations are assumed to compete to explain evidence in a zero-sum game.

Evidence evaluation and the likelihood ratio

The likelihood ratio (LR) – the probability of an item of evidence given the hypothesis is true, divided by the probability of that same evidence given the hypothesis is false – is used to determine the probative value of evidence in legal, forensic, medical, and other domains of reasoning under uncertainty (Finklestein, 2009; Fenton & Neil, 2012). Evidence is considered probative if the LR is greater than 1: that is, when the evidence is more likely if the hypothesis is true rather than false. Thus, if evidence is equally likely to occur whether the hypothesis is true or false, the LR equals 1, and the evidence is considered non-probative.

However, as in the Barry George case, the LR can also be misapplied, with deleterious consequences. An LR equal to 1 only implies that evidence is non-probative if the hypotheses that make up the ratio are mutually exclusive and exhaustive (typically a target hypothesis and its negation - 'Barry George fired the gun' vs 'Barry George *did not* fire the gun'). Crucially, when the target hypothesis and an alternative hypothesis that is not the negation of the target are under consideration (e.g. whether Barry George fired the gun, and whether the police mishandled the evidence), assumptions of mutual exclusivity and exhaustiveness are often not met (i.e. both or neither hypothesis may be true). As a consequence, even if the likelihood ratio is equal to 1, it is a mistake to infer that the evidence is not probative of the target hypothesis (see Fenton et al., 2014). This mistake can arise in any domain where evidence has multiple independent explanations, the general case for which is illustrated by the common-effect Bayes Net structure of Fig. 1.



Test	¬H1,¬H2	¬H1,H2	H1,¬H2	H1,H2
E	0.01	0.9	0.9	0.99
¬E	0.99	0.1	0.1	0.01

Fig. 1.

Common-effect scenario, with H1 and H2 representing the two claims, both candidate causes of positive (true) test results; the conditional probability table of test is included, where P(E|H1, H2) results from assuming a noisyOR, see Pearl, 1988). Priors of both claims are arbitrarily set to 0.5. **t_0:** No evidence has been observed. **t_1:** Evidence observed as True.

In this example the evidence of a positive test is observed (t_0 to t_1), increasing the probability of *both* hypotheses – despite the fact that the LR of the evidence for H1 against H2 is equal to 1. Equivalent examples include multiple diseases and a medical test, multiple explanations of a person's behaviour (Nisbett & Ross, 1980), or multiple explanations for a crime. In all such cases, the danger is that people mistakenly judge crucial evidence to be non-probative, because they focus on whether the evidence discriminates between the target hypothesis and an alternative hypothesis (H1 versus H2), rather than between the target hypothesis and its negation (H1 versus ~H1). It is the latter comparison that is critical for determining whether the evidence supports (or undermines) the target hypothesis (H1).[1]

In our example we used priors of 0.5 for each hypothesis, but this was an illustrative choice. In fact, the key pattern of inference - whereby evidence that has LR=1 for H1 vs H2 is still probative for H1 vs not-H1 – holds irrespective of the priors of the hypotheses (so long as these are neither 0 or 1), given plausible assumptions about the conditional probability table for the evidence E (see proofs in Supplementary materials A).

Zero-sum reasoning

We posit that this error is based on the misconception of evidential support as a finite, shared resource across the hypotheses under contention. This "zero-sum" conceptualisation of support is appropriate only *if* hypotheses truly are both exclusive and exhaustive. But, in general, evidential support is not a zero-sum game, and reasoning from this assumption can lead to ignoring valuable evidence.

The notion of zero-sum effects has been explored in psychology, where people inappropriately "cap" available resources – whether predictions of student grade quality (Meegan, 2010) or "fixed-pie" beliefs (Smithson & Shou, 2016) – with the resulting assumption that positivity in one domain corresponds to negativity in another (e.g. "When the rich get richer, the poor get poorer."). This effect relates to the notion of "hydraulic" action (attribution to one must be balanced by substitution from another), explored in work on social attribution (Kanouse, 1972; Lepper & Greene, 1978; Nisbett & Ross, 1980), where people judge that more support for a behaviour (e.g. an angry outburst) due to an intrinsic explanation (e.g. being an angry person) must correspond to less support for an extrinsic explanation (e.g. the situation). We propose a

[1] Here we focus on the qualitative notion of evidential support, whereby a hypothesis H is supported by evidence E if P(H|E) > P(H), which by Bayes's rule is equivalent to P(E|H) > P(E). This notion is uncontroversial, but leaves open the question of the appropriate quantitative measure of degree of evidential support (Crupi et al, 2007). The latter question does not impact on the arguments in this paper, which require only the qualitative notion.

zero-sum reasoning fallacy, wherein the degree of support across multiple explanations is considered fixed, such that evidence that does not distinguish between these explanations is deemed irrelevant. Critically, this is based on a false assumption of exclusivity and exhaustiveness across explanations, when in fact the same evidence can offer support for *both* explanations.

Experiment 1

Experiment 1 demonstrates the zero-sum fallacy. We predict that when presented with evidence that should increase support for both hypotheses, lay reasoners will erroneously judge it irrelevant. Conversely, when presented with evidence that should decrease support for both hypotheses, we predict reasoners will correctly use this evidence to disconfirm both hypotheses, because correct responding ("ruling out" explanations) does not require hypotheses to be treated as non-exclusive.

Method

Participants. A total sample size of 50, with 25 participants per test result condition, (yielding 100 observations) was predetermined. Participants were recruited and participated online through MTurk (https://www.mturk.com/). Those eligible for participation had a 95% and above approval rating from over 100 prior tasks. Participants were English speakers, located in the United States. One participant was removed for incomplete responses. Of the 49 participants remaining, 26 were female. The mean age was 33.37 (*SD* = 10.27). Participants were paid $1 for their time (*Median* = 5.87 minutes, *SD* = 5.54).

Materials and Procedure. Participants all completed basic demographics (age, gender, native language) before moving onto the scenarios. Each participant completed the four scenarios (see Supplementary Materials B) in a random order. Participants were assigned to one of two conditions: Either all scenarios contained positive test results, or all contained negative test results. In all cases, there was both a target and alternative explanation for the test result (in line with the structure of Fig.1). For each scenario, participants were asked to make a judgment of "Yes", "No" or "Cannot Tell" when posed with the following example format (negative test result condition in braces):

"*Does a positive [negative] Griess test result give any support to the claim that Ann has [not] handled explosives?*"

On a separate page, after each scenario judgment, participants were asked to "Please briefly provide some reasoning for your decision regarding the previous scenario in the text box below." (Not reported in the present paper). Along with demographics, scenario order and time taken were recorded. Participants were paid for their time.

Results

All analyses were Bayesian, and performed using the JASP statistical software (JASP Team, 2016)[2]. Importantly, the use of Bayes Factors allows us to infer evidence for the null hypothesis, wherein $BF_{10} < 1/3^{rd}$ is considered strong support for the null (Dienes, 2014).

Judgment Data. Each of the 49 participants made 4 judgments. Fig. 2 shows the mean proportions of these judgments, split by test result condition.

[2] All Bayesian analyses use an objective (uninformed) prior.

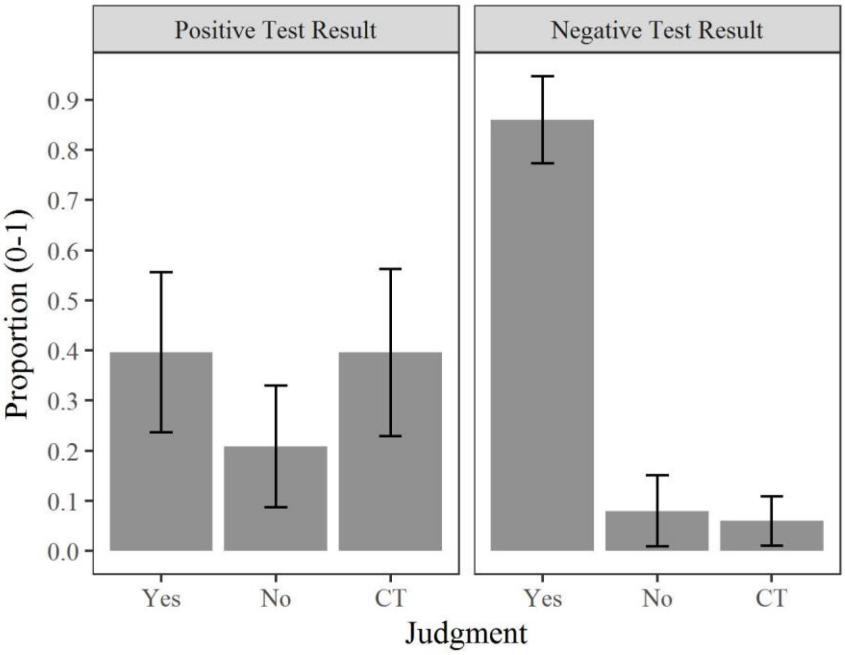


Fig. 2.

Experiment 1. Mean proportions of judgments (CT = “Cannot Tell” responses), split by test result condition. Error bars reflect 95% Confidence Intervals.

To analyse the data, for each participant judgments were coded as either correct (1; “Yes”) or incorrect (0; “No” or “Cannot Tell”). These responses were then summed across the 4 scenarios, resulting in a single summary variable (“Sum Correct”) for each participant, bound between 0 and 4. Consequently, using a Bayesian independent samples T-test, participants in the positive test result condition (*M* = 1.58, *SD* = 1.38, *N* = 24) made significantly fewer correct responses than participants in the negative test result condition (*M* = 3.44, *SD* = 0.79, *N* = 25), BF_{10} = 26463.93, δ = 1.563 (95% CI[3]: [0.903, 2.225]). Finally, correct responding was compared to chance level (test value = 1.33). Whilst correct responding was significantly greater than chance for participants in the negative test result condition, $BF_{10} = 1.03 * 10^{10}$, δ = 2.656 (95% CI: [2.003, 3.407]), participants in the positive test result condition showed strong evidence for the null (i.e. were at chance level), BF_{10} = 0.31, δ = 0.162 (95% CI: [-0.208, 0.540]).

[3] CI here refers to credibility interval. δ refers to Cohen’s d effect size.

Lastly, Bayesian contingency tables were used to check whether scenario order or type influenced judgments. Neither scenario order, $BF_{10} = 3.56 * 10^{-4}$, nor type, $BF_{10} = 0.002$, influenced judgments, with decisive evidence for the *null* in both instances.

Discussion

Positive evidence is judged as irrelevant significantly more than negative evidence. This fits with our predictions, given the negative test does not require the introduction of "new" resources (the "sum" part of zero-sum), but instead reduces support (i.e. the negative test *disconfirms* both hypotheses). These results are not influenced by scenario order (i.e. no effects of learning or attentional attrition), or type (indicating context generalizability).

Experiment 2

Experiment 2 examines two key questions. First, is the error due to a failure to consider that the hypotheses are non-exhaustive? Second, are "Cannot Tell" responses (which we consider erroneous) due to low confidence rather than a genuine misinterpretation of the value of the positive test result?

Method

Participants. Participants were recruited using the same protocol as in Experiment 1. A sample size of 200 was predetermined, based on a conservative estimate for a possible interaction between test result and exhaustiveness intervention (see below) factors. Of the 200 participants recruited (50 per group, see below), three were removed whose native language was not English, and four further participants were removed for incomplete responses. Of the 193 participants remaining, 88 were female. The mean age was 36.27 (*SD* = 10.93). Participants were paid $1 for their time (*Median* = 7.37 minutes, *SD* = 5.54).

Materials and Procedure. The materials used were identical to those of Experiment 1, with the same general procedure, but to address the questions of Experiment 2, the following changes were made:

To address the exhaustiveness issue, a between-subject factor was introduced, in which an explicit statement regarding non-exhaustiveness was either present ("Non-Exhaustiveness Statement") or absent (control). This, in conjunction with the test result between-subject manipulation, led to a 2 x 2 design. The non-exhaustiveness statement preceded the standard judgment question, and used the following structure:

"*Please note, it is possible that [Subject]* ***neither*** *[H_1]* ***nor*** *[H_2].*"

To address the confidence question, a confidence measure was included directly below the judgment question. The phrasing of this question was "How **confident** are you that your **response is correct**?" using a slider to indicate from 0% to 100% (no default value; see Supplementary Materials C for an example scenario).

Accordingly, as participants completed each of the 4 scenarios in a random order, they made a judgment, expressed their confidence in that judgment, before moving on to provide some reasoning.

Results

Judgment Data. Each of the 193 participants made 4 judgments, resulting in a total of 772 judgments. Fig. 3 shows the mean proportions of these judgments, split by test result condition (columns) and exhaustiveness manipulation (rows).

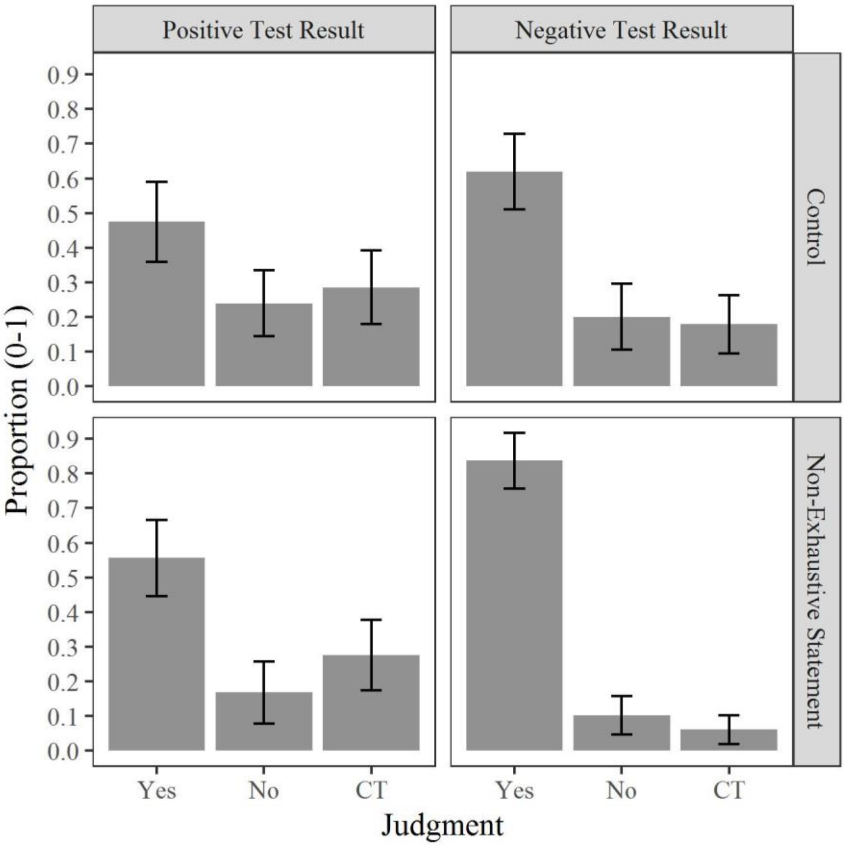


Fig. 3.

Experiment 2. Mean proportions of judgments (CT = "Cannot Tell" responses), split by test result condition (columns) and exhaustiveness manipulation (rows). Error bars reflect 95% Confidence Intervals.

As in Experiment 1, participant judgments were coded into a single, summary correct responding variable. We conducted a Bayesian ANOVA with the test result and exhaustiveness manipulation between subject factors. As can be seen in Table 1, correct responding was significantly higher in negative (vs positive) test result conditions, $BF_{Inclusion}$[4] = 607.57, and the non-exhaustiveness statement (vs control) also led to higher correct responding, $BF_{Inclusion}$ = 9.02. As the interaction was not significant, the model with these two main factors was considered the best fit, BF_M = 6.86, and significant overall, BF_{10} = 5031.76. Breaking down the main effect of

[4] $BF_{Inclusion}$ is the change in odds from the sum of prior probabilities of models including the effect to the sum of posterior probabilities of models including the effect.

the exhaustiveness manipulation by test result, the increase in correct responding was found in the negative test result condition (*N* = 95), BF_{10} = 29.37, δ = -0.64 (95% CI: [-1.046, -0.247]), but not in the positive test result condition (*N* = 98), BF_{10} = 0.36, δ = -0.192 (95% CI: [-0.57, 0.177]).

Table 1.

Experiment 2: Correct responding descriptives and chance responding analysis, split by condition.

Test Result	Exhaustiveness Statement	*M*	*SD*	*N*	*≠1.33 (BF_{10})*	*δ*	*δ 95% CI*
Negative	Control	2.48	1.41	46	10813	0.777	0.462, 1.112
	Non-exhaustiveness	3.35	1.07	49	$3.883 * 10^{14}$	1.847	1.455, 2.304
Positive	Control	1.90	1.54	49	2.988†	0.348	0.066, 0.639
	Non-exhaustiveness	2.22	1.46	49	260.2	0.583	0.279, 0.886

Note: *† = anecdotal evidence.*

As can be seen in the final column of Table 1, all correct responding rates were significantly greater than chance level, with the single exception of the positive test result participants who did not receive the non-exhaustiveness statement.

Lastly, using Bayesian contingency tables, the potential confounds of scenario order and type did not impact judgments, with strong support for the null in both the former, BF_{10} = $8.93 * 10^{-6}$, and the latter, BF_{10} = 0.02.

Confidence Data. Fig. 4 shows the boxplot breakdown of confidence by judgment type (within-pane), test result condition (columns) and exhaustiveness manipulation (rows).

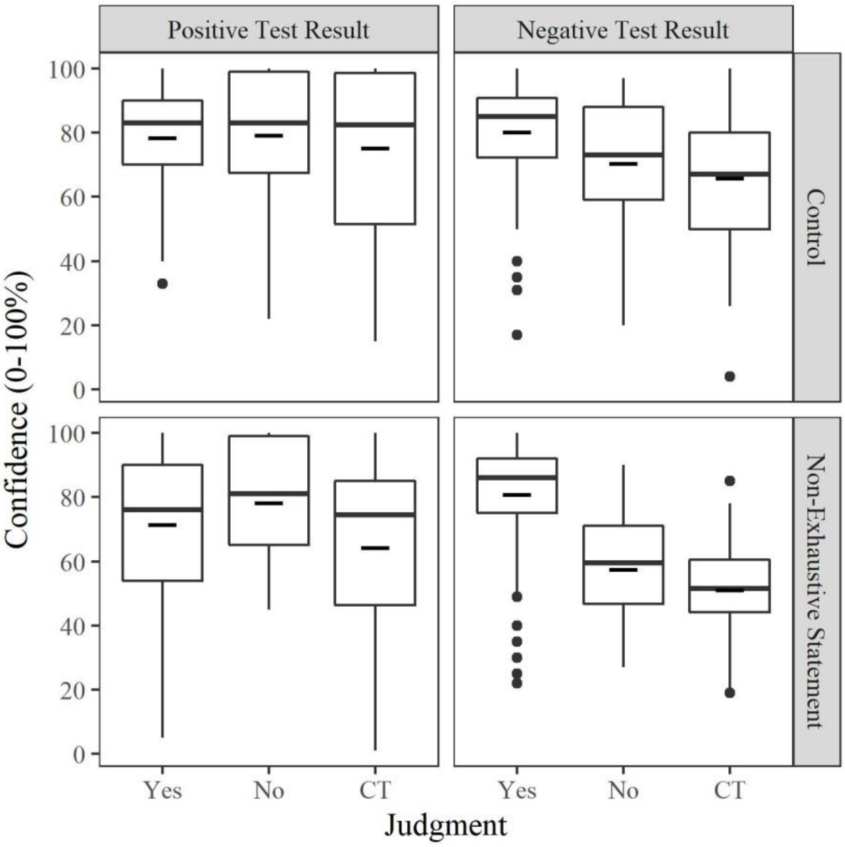


Fig. 4.

Experiment 2. Confidence in judgments (CT = "Cannot Tell" responses), split by test result condition (columns) and exhaustiveness manipulation (rows). Whiskers represent +/- 1.5 * IQR. Means shown as short crossbars.

A Bayesian ANOVA was run on the three variables of interest, Judgment (3) x Test Result Condition (2) x Exhaustiveness Manipulation (2). Hierarchical model comparisons revealed significant main effect of judgment on confidence, $BF_{Inclusion} = 1.18 * 10^8$, with "Cannot Tell" responses as least confident, and "Yes" responses as most confident. Positive test result conditions also led to higher confidence, $BF_{Inclusion} = 1434.77$, whilst the exhaustiveness manipulation significantly decreased confidence, $BF_{Inclusion} = 8.41$. Lastly, the analysis yielded a significant interaction between test result condition and judgments, $BF_{Inclusion} = 7510.11$, with the model including this interaction term yielding the most significant model improvement, $BF_M = 18.09$, and decisive evidence overall, $BF_{10} = 3.63 * 10^8$. To explore this interaction further, two further ANOVA were performed on both positive and negative test result conditions in isolation. This revealed that confidence was not significantly affected by judgment type in the positive test result condition (see left-hand column of Fig. 4; *N* = 98), $BF_{10} = 0.81$, whilst those in the negative test condition were *decisively* less confident in “Cannot Tell” and “No” judgments than “Yes” judgments (see right-hand column of Fig. 4; *N* = 95), $BF_{10} = 5.55 * 10^{23}$.

Discussion

Experiment 2 replicates Experiment 1, as positive evidence is once again judged as irrelevant significantly more than negative evidence. Although the non-exhaustiveness statement was effective in improving judgments, primarily applied to negative evidence (although correct responding in positive evidence was above chance level, suggestive of a weak effect). Crucially, confidence estimates reveal that erroneous judgments for positive evidence were made as confidently as correct responses (providing evidence against a “low confidence bin” explanation). Finally, judgments again were unaffected by either scenario order or scenario type.

Experiment 3

Experiment 3 explored the impact of exclusivity by manipulating whether or not participants were explicitly told that the hypotheses were mutually exclusive. Further, it examined whether the zero-sum fallacy still holds when the “leak” value of the test (i.e. the probability of a false positive when neither hypothesis is true), is given. This allows us to rule out the possibility that “no” or “cannot tell” judgments in the positive test condition are due to an assumption that the test is *generally* undiagnostic. More precisely, reasoners may assume that as multiple hypotheses can entail a positive result, the test generally yields positive results (irrespective of hypotheses), making the test worthless. This would be represented by an inflated leak value (e.g. 90%), and so specifying a low leak value (e.g. 1%) rules out this explanation.

Method

Participants. Participants were recruited using the same protocol as in Experiment 1. A sample size of 207 was predetermined, based on the rationale of Experiment 2, taking into account previous rates of ineligible participants / incomplete data submissions. Accordingly, of the 207 recruited, 6 were removed whose native language was not English (2), were living outside the US (3), or submitted incomplete data (1). Of the 201 participants remaining, 112 were female. The mean age was 37.51 (*SD* = 12.31). Participants were paid $1 for their time (*Median* = 7.56 minutes, *SD* = 6.57).

Materials and Procedure. The materials and general procedure generally followed that of Experiment 2, barring the following exceptions:

Across all conditions, a statement was included to indicate the probability of a false positive (i.e. a test coming back positive when neither hypothesis was true). This took the general form of:

"If neither [H1] nor [H2] is true, there is only a [X%] chance of the test being positive."

The specific wording and value of the false positive was tailored to each scenario (though the latter was fixed between 0.5% and 3%; details of which can be found in Supplementary Materials C).

Along with the between-subject factor of test result condition (positive or negative; common to Experiments 1 and 2), an additional, 2-level between-subject exclusivity manipulation was added (present or absent; making a 2x2 between-subject design). This manipulation consisted of either the presence or absence of an explicit exclusivity constraint across all scenarios. This constraint took the general form of:

"Importantly, [CONSTRAINT] means it is not possible for **both** to be true (i.e. [Subject] can't have/be [H1] **and** [H2])."

Specific wording of these exclusivity constraint manipulations for each scenario are also provided in Supplementary Materials D. As an example (taken from the Brain Tumor scenario):

“Importantly, Gary's other symptoms mean it is not possible for **both** to be true (i.e. Gary can't have a tumor **and** EOD).”

Accordingly, as participants completed the 4 scenarios in a random order, they made a judgment, expressed their confidence in that judgment, and then provided some reasoning.

Results

Judgment Data. Each of the 201 participants made 4 judgments, resulting in a total of 804 judgments. Fig. 5 shows the mean proportions of these judgments, split by test result condition (columns) and exclusivity manipulation (rows).

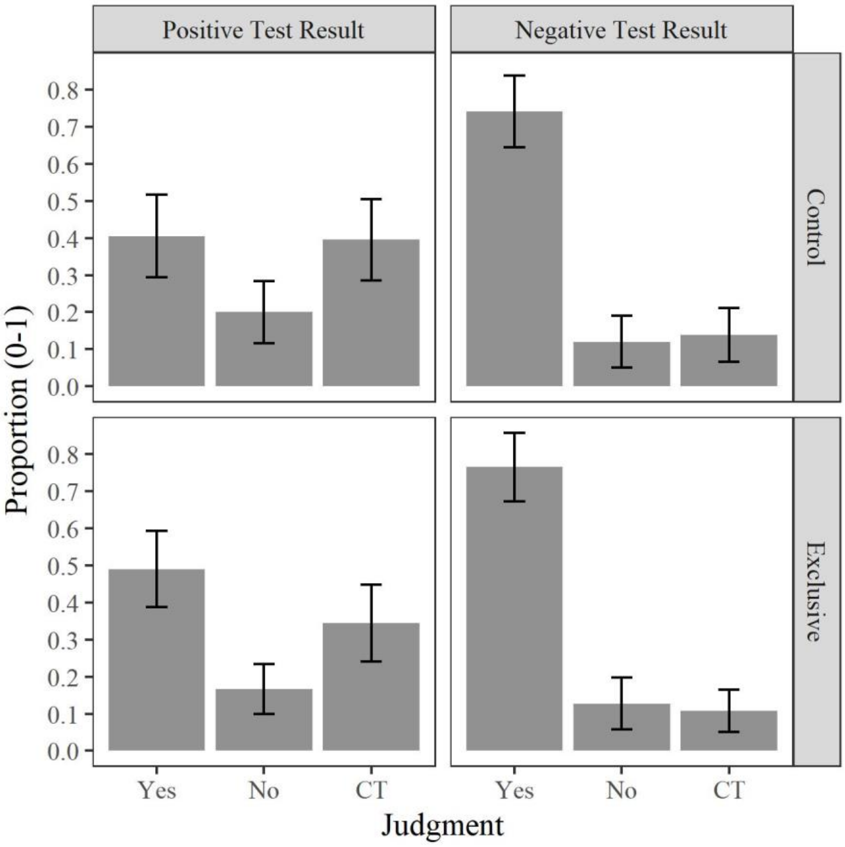


Fig. 5.

Experiment 3. Mean proportions of judgments (CT = "Cannot Tell" responses), split by test result condition (columns) and exclusivity manipulation (rows). Error bars reflect 95% Confidence Intervals.

Following the analysis protocol of the preceding experiments, participant judgments were again coded into a single, summary correct responding variable, which was used as the dependent variable in subsequent Bayesian ANOVA. Hierarchical model comparison found that whilst correct responding was significantly higher in negative (vs positive) test result conditions (right vs left columns of Fig. 5), $BF_{Inclusion} = 5.74 * 10^6$, there was strong evidence for a null effect of exclusivity manipulation, $BF_{Inclusion} = 0.281$. Consequently, the model with only a main effect of test result was both the best fit, $BF_M = 9.487$, and significant overall, $BF_{10} = 4.045 * 10^6$. As in Experiment 2, the main effect of the exclusivity manipulation was broken down by test result. In line with the overall analysis, there was no effect of the exclusivity manipulation in either the positive (*N* = 98), $BF_{10} = 0.391$, $\delta = 0.203$ (95% CI: [-0.16, 0.592]), or negative (*N* = 103), $BF_{10} = 0.222$, $\delta = 0.067$ (95% CI: [-0.305, 0.426]), test result conditions.

Table 2.

Experiment 3: Correct responding descriptives and chance responding analysis, split by condition.

Test Result	Exclusivity Manipulation	*M*	*SD*	*N*	*≠1.33 (BF_{10})*	*δ*	*δ 95% CI*
Negative	Exclusive	3.06	1.26	51	$2.815 * 10^{10}$	1.34	0.969, 1.715
	Control	2.96	1.33	52	$1.332 * 10^9$	1.194	0.877, 1.562
Positive	Exclusive	1.96	1.35	48	13.60	0.434	0.144, 0.736
	Control	1.62	1.50	50	0.37†	0.18	-0.088, 0.456

Note: *† = anecdotal evidence*

In line with previous experiments, correct responding in negative test result conditions was significantly greater than chance (top 2 rows of Table 2), which occurred irrespective of exclusivity manipulation. Interestingly, in the positive test result condition, when an exclusivity constraint was made explicit, correct responding was significantly greater than chance level – which was not the case in the control condition.

Lastly, as in Experiments 1 and 2, using Bayesian contingency tables, judgments were shown to be unaffected by scenario order (N = 804), $BF_{10} = 7.465 * 10^{-5}$, and scenario type ($N$ = 804), $BF_{10} = 2.177 * 10^{-4}$, with very strong evidence for the null in both cases.

Confidence Data. Fig. 6 shows the boxplot breakdown of confidence by judgment type (within-pane), test result condition (columns) and exclusivity manipulation (rows).

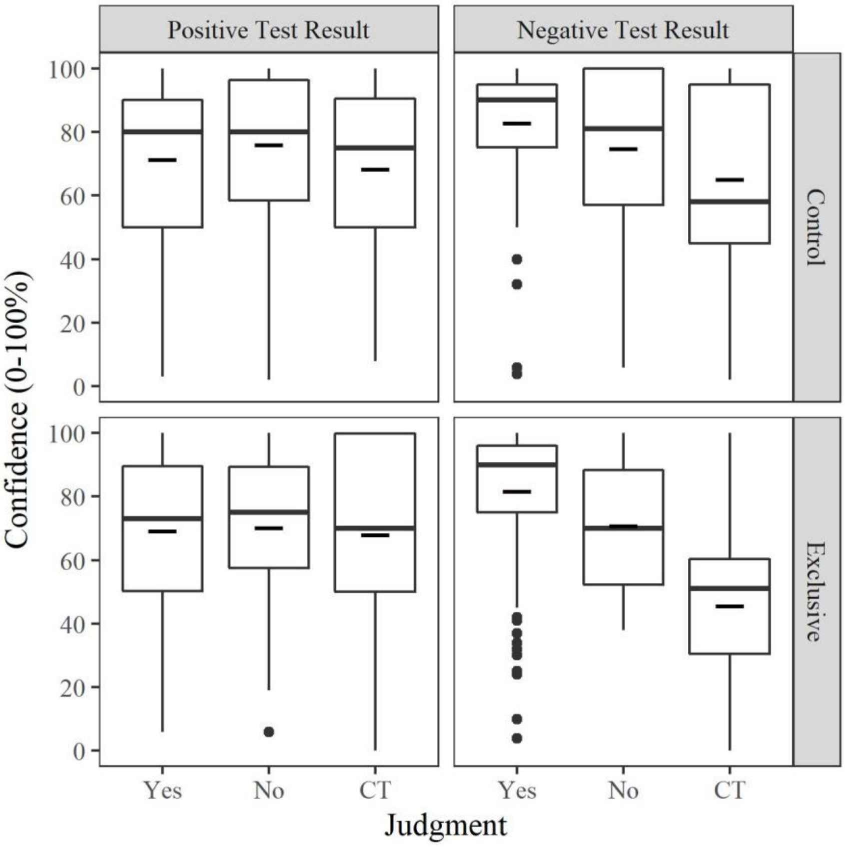


Fig. 6.

Experiment 3. Confidence in judgments (CT = "Cannot Tell" responses), split by test result condition (columns) and exclusivity manipulation (rows). Whiskers represent +/- 1.5 * IQR. Means shown as short crossbars.

A Bayesian ANOVA was run on the three variables of interest, Judgment (3) x Test Result Condition (2) x Exclusivity Manipulation (2). Hierarchical model comparisons revealed a significant main effect of judgment on confidence, $BF_{Inclusion} = 2.07 * 10^8$, with “Cannot Tell” responses as least confident, and “Yes” responses as most confident. There was also a main effect of test result condition, $BF_{Inclusion} = 132260.63$, with judgments in negative test result conditions higher than positive, but strong evidence for a null effect of exclusivity manipulation, $BF_{Inclusion} = 0.175$. Lastly, the analysis yielded a significant interaction between test result condition and judgments, $BF_{Inclusion} = 132899.66$, with the model including these significant terms yielding the most significant model improvement, $BF_M = 36.78$, and decisive evidence overall, $BF_{10} = 1.562 * 10^{12}$. To explore this interaction further, a second round of ANOVA were performed on both positive and negative test result conditions in isolation. This revealed that confidence was not significantly affected by judgment in the positive test result condition (see left-hand column of Fig. 6; $N = 392$), $BF_{10} = 0.079$, whilst those in the negative test result condition were *decisively* less confident in “Cannot Tell” and “No” judgments than “Yes” judgments (see right-hand column of Fig. 6; $N = 412$), $BF_{10} = 5.266 * 10^{11}$.

Discussion

Experiment 3 demonstrates that the distinctive zero-sum pattern of reasoning holds even when leak values for tests are included, and when exclusivity constraints are made explicit. Replicating Experiment 2 these errors are given with equally high confidence as correct responses, corroborating a “misplaced faith” in such errors.

General Discussion

Three experiments present evidence for the zero-sum fallacy. Experiment 1 showed the fallacy in positive test cases, comparing it directly with negative test cases, where no such error is made. Experiment 2 explicitly stated that the candidate hypotheses were non-exhaustive; an intervention that reduced errors in negative test cases (although we note some weak evidence for improved correct-responding in positive test cases). Experiment 3 showed no significant impact on the pattern of reasoning when hypotheses were stated to be exclusive, and also that erroneous reasoning was not due to participants believing that the tests were generally non-diagnostic.

Further experiments have also shown that the fallacy holds even when likelihoods differ.[5] In addition, the inclusion of confidence measures showed that both erroneous and correct judgments were held with high confidence.

We conjecture that this bias arises because people treat evidence as a zero-sum game, whereby alternative hypotheses compete for evidential support. Thus, evidence that favours one hypothesis must thereby disfavour alternative hypotheses. This assumption prohibits people from seeing that the same piece of evidence can simultaneously confirm alternative hypotheses. More precisely, lay reasoners assume that evidence which is equally predicted by two competing hypotheses offers no support for either hypothesis. However, this assumption only holds when the competing hypotheses are mutually exclusive and exhaustive. In the contexts presented in these experiments, and in many real-world contexts such as law and medicine, these conditions do not hold, and yet people persist in disregarding evidence that is genuinely probative of the key hypothesis of interest.

Reasoning under zero-sum assumptions seems to be a compelling heuristic that will often simplify inference and promote clear-cut decision making. But when conditions of exclusivity or exhaustiveness fail, as in many real world situations, reasoners will overlook crucial evidence.

[5] An additional experiment that looked at differing likelihood values for $P(E|H_1)$ and $P(E|H_2)$, found no impact on the effects described here. Full details are included in Supplementary Materials E.

AUTHOR CONTRIBUTIONS

T. D. Pilditch, D. Lagnado, and N. Fenton developed the study concept. All authors contributed to the study design. Testing and data collection were performed by T. D. Pilditch. T. D. Pilditch performed the data analysis and interpretation under the supervision of D. Lagnado. T. D. Pilditch drafted the manuscript, and D. Lagnado and N. Fenton provided critical revisions. N. Fenton provided the mathematical proofs in the appendices. All authors approved the final version of the manuscript for submission.

ACKNOWLEDGEMENTS

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), under Contract [2017-16122000003]. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

OPEN PRACTICES

All data and materials have been made publicly available via the Open Science Framework at (https://osf.io/wnu9f/).

REFERENCES

Crupi, V., Tentori, K., & Gonzalez, M. (2007). On Bayesian measures of evidential support: Theoretical and empirical issues. *Philosophy of Science*, *74*(2), 229-252.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in psychology*, 5, 1-17.

Fenton, N., Berger, D., Lagnado, D., Neil, M., & Hsu, A. (2014). When 'neutral' evidence still has probative value (with implications from the Barry George Case). *Science and justice*, *54*(4), 274-287.

Fenton, N., & Neil, M. (2012). *Risk assessment and decision analysis with Bayesian networks*. Crc Press.

Finkelstein, M. O. (2009). Probability. In *Basic Concepts of Probability and Statistics in the Law* (pp. 1-18). Springer, New York, NY.

JASP Team. (2016). JASP (Version 0.8.0.0).

Kanouse, D. E. (1972). Language, labelling, and attribution. In E. E. Jones (Ed.), *Attribution: Perceiving the causes of behaviour* (pp. 121–135). Morristown, NJ: General Learning Press.

Lepper, M. R., & Greene, D. (1978). *The hidden costs of reward*. Hillsdale, NJ: Elbaum.

Meegan, D. V. (2010). Zero-sum bias: perceived competition despite unlimited resources. *Frontiers in psychology*, *1*, 191.

Nisbett, R., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgement*. Englewood Cliffs, NJ: Prentice-Hall.

Pearl, J. (1988) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, Calif: Morgan Kaufmann Publishers, Inc.

Smithson, M., & Shou, Y. (2016). Asymmetries in responses to attitude statements: The example of "zero-sum" beliefs. *Frontiers in psychology*, *7*, 984.