Best, Second-Best, and Good-Enough Explanations: How They Matter to Reasoning.

Igor Douven, Patricia Mirabile

Best, Second-best, and Good-enough Explanations: How They Matter to Reasoning*

Igor Douven† SND/CNRS/Sorbonne University

Patricia Mirabile SND/Sorbonne University

*All Supplementary Information as well as all data and the R script used for the analyses can be downloaded from http://doi.org/10.17605/OSF.IO/CJB9U.

†Corresponding author: 1 rue Victor Cousin, 75005 Paris, France, igor.douven@paris-sorbonne.fr.

Abstract

There is a wealth of evidence that people's reasoning is influenced by explanatory considerations. Little is known, however, about the exact form this influence takes, for instance about whether the influence is unsystematic or due to people's following some rule. Three experiments investigate the descriptive adequacy of a precise proposal to be found in the philosophical literature, to wit, that we should infer to the best explanation, provided certain additional conditions are met. The first experiment studies the relation between the quality of an explanation and people's willingness to infer that explanation when only one candidate explanation is given. The second experiment presents participants always with two explanations and investigates the effect of the presence of an alternative on the participants' willingness to infer the target explanation. While Experiments 1 and 2 manipulate explanation quality and willingness to infer to the best explanation *between* participants, Experiment 3 manipulates those measures *within* participants, thereby allowing to study the influence of explanatory considerations on inference at the individual level. The third experiment also studies the connection between explanation quality, willingness to infer, and metacognitive confidence in the decision to infer. The main conclusions that can be drawn from these experiments are that (i) the quality of an explanation is a good predictor of people's willingness to accept that explanation, and a better predictor than the prior probability of the explanation, and (ii) if more than one possible explanation is given, people are the less willing to infer the best explanation the better they deem the second-best explanation.

Keywords: explanation; inference; metacognitive confidence; probability; reasoning.

Sometimes the evidence leaves on the table only one reasonable explanation for a given fact. For instance, human-induced climate change is at the moment the only reasonable explanation for the rising sea levels and a host of other phenomena. Most agree that in this type of situation we can safely infer the truth of that explanation (even though the rising sea levels together with the other evidence in favor of the hypothesis of human-induced climate change do not logically *entail* that hypothesis).

However, often in science as well as in everyday life, the situation is not so clear-cut, and we are faced with multiple potential explanations for a phenomenon of interest. If two or more of those candidate explanations look equally plausible in the light of all the evidence, then we may have to suspend judgment about which of them is true (if any of them *is* true, given that there is always the possibility that we are overlooking further potential explanations). But what are we licensed to infer if one of the candidate explanations stands out compared to its rivals and offers a better explanation than they do?

According to some philosophers of science, such a status warrants the—possibly qualified—acceptance of the explanation. That is to say, its status qua best explanation licenses us to believe it, or to believe it to a high degree, or to believe that it is closer to the truth than any other candidate explanation (to mention just a few of the possible qualifications that have been discussed in the literature).

This type of inference is known as "Inference to the Best Explanation" (IBE), although—as the foregoing already suggests—there is no unanimity on what exactly the inference amounts to. Common textbook versions of IBE have it that we are licensed to infer the truth of the best explanation of our evidence (see, e.g., Vogel, 1998, or Psillos, 2004), but many philosophers regard formulations to this effect as too crude, for a variety of reasons.[1]

[1] IBE, in any of the versions discussed in the philosophical literature, is to be distinguished from what Wilkenfeld and Lombrozo (2015) have dubbed "Explaining for the Best Inference" (EBI). IBE and EBI both point at close connections between explanation and inference, but while IBE sees explanatory considerations as warranting inference (under certain conditions), EBI emphasizes the importance of the activity of explanatory reasoning for improving our epistemic standing, and thereby improving the inferences we make. As Wilkenfeld and Lombrozo note, IBE and

For one, that a given candidate explanation is better than all its rivals does not by itself mean that it is a good explanation, absolutely speaking. And we would not really want to infer the truth of a poor explanation of the evidence, even if the competing explanations are poorer still. This point was made forcefully by Lipton (1993, 2004), who proposed that we infer the truth of the best explanation only if the best is good enough.

For another, even if there is a best explanation for the evidence at hand and that best explanation is also good enough, it is still an open question as to how much better an explanation it is than its closest competitor(s). The answer to this question may make a difference in what we can infer from the evidence. Specifically, Bird (2010, p. 346) suggests that, given a best and good-enough explanation, we are entitled to infer its truth only if it is "significantly better than its nearest rival." If the second-best explanation offers a perfectly good explanation for the evidence as well and is just barely topped by the best explanation, then, Bird says, "our faith in that slightly better one must be slim" (*ibid.*).

Philosophers have discussed the relationship between explanation and inference from a purely normative standpoint; they have not been concerned with the question of whether their proposals are descriptively adequate. Meanwhile, psychologists have shown that explanation is central to various cognitive processes, including categorization (e.g., Williams & Lombrozo, 2010), generalization (e.g., Lombrozo & Gwynne, 2014), learning (e.g., Baillargeon & DeJong, 2017; Lombrozo, 2016; Rittle-Johnson & Loehr, 2017; Sidney, Hattikudur, & Alibali, 2015; Williams & Lombrozo, 2013), understanding (e.g., Keil, 2006; Walker & Lombrozo, 2017), and semantic and pragmatic processing (Bunt & Black, 2000; Douven, 2016a; Douven, Elqayam, Singmann, & van Wijngaarden-Huitink, 2018; Douven & Verbrugge, 2010; Hobbs, 1992, 2004). There is also empirical work related to IBE, but for the most part this concerns probabilistic versions of IBE, which relate explanatory considerations to the updating of subjective probabilities. Specifically, such versions are like Bayes' rule, except that they take into account the explanatory power of whichever hypotheses are at issue and can give an extra probability boost to the best-

EBI can peacefully coexist.

explaining hypothesis. Douven and Schupbach (2015a, 2015b) report evidence that, in certain contexts, such versions of IBE predict people's probability updates more accurately than does Bayes' rule (see also Bes, Sloman, Lucas, & Raufaste, 2012).[2] However, there is so far almost no empirical work directly concerning more standard versions of IBE, which relate explanation to the categorical acceptance of hypotheses. In this paper, we focus on precisely that relation.

In doing so, we shall be especially interested in the following research questions:

Q1. Is the quality of an explanation a good predictor of people's willingness to accept that explanation?

Q2. Given a potential explanation of some phenomenon, does it make any difference, in regard to the perceived quality of that explanation and in regard to people's willingness to accept it, whether or not a second explanation is introduced (and if so, why)?

Q3. Given two rival explanations of some phenomenon, does the *magnitude* of their difference in quality make a difference to people's willingness to accept the better of the two?

We report three experiments that are meant to shed light on these questions. The first experiment, focusing most directly on Q1, investigates the relation between the quality of an explanation and people's willingness to infer that explanation when it is the only explanation given. The second experiment presents participants always with two explanations and, by comparing the results with those from Experiment 1, investigates the effect of the presence of an alternative on participants' willingness to infer the target explanation. Experiment 3 investigates the same effect, but while Experiments 1 and 2 manipulate explanation quality and willingness to infer to the best explanation *between* participants, Experiment 3 manipulates those measures *within* participants. As a result, the data from the third experiment permit us to study the influence of explanatory considerations on inference at the individual level. Furthermore, Experiment 3 also relates the link between explanation and inference to the issue of metacognition, an issue

[2] Bayes' rule has independently been shown to sometimes be at stark variance with people's actual updating practices; see, e.g., Baratgin and Politzer (2007), Fischhoff and Lichenstein (1978), Robinson and Hastie (1985), Schum and Martin (1982), and Tversky and Kahneman (1974).

that will be seen to suggest a possible answer to a question that arises from the results of the second experiment.

Theoretical background

In the 1970s and 1980s, philosophers considered IBE to be a matter of course (Boyd, 1981; McMullin, 1992; Putnam, 1975). The rule came under a cloud, however, with the advent of the Bayesian paradigm. According to Bayesians, there is only one rule for rational belief change, and that is Bayes' rule (and generalizations thereof, such as Jeffrey conditionalization and entropy minimization). Bayesians have offered Dutch book arguments (e.g., de Finetti, 1937; Ramsey, 1926; Teller, 1973) as well as inaccuracy-minimization arguments (Joyce, 1998; Rosenkrantz, 1992) in defense of their position, but both types of arguments are questionable (Douven, 1999, 2013, 2016b, 2017b; Douven & Wenmackers, 2017).

The Bayesian community in psychology (e.g., Oaksford & Chater, 2007; Over, 2009) has from the start been much more open-minded on the issue of belief change than Bayesian philosophers of science tend to be. Oaksford and Chater (2013, p. 374) are very explicit about this issue when they state that "it is unclear what are the rational probabilistic constraints on dynamic inference."

Also note that maintaining that people tend to infer to the best explanation is not to deny that reasoning is fundamentally probabilistic. In one of two major approaches to causality (Pearl, 1988, 2000; Sloman, 2005), this notion is analyzed by means of graphical models, which are at the same time statistical models, encoding relations of probabilistic dependence and independence.[3] In recent years, these so-called causal Bayes nets have gained great popularity in psychology, where they have helped to illuminate various types of reasoning involving causality, including diagnostic reasoning (Meder & Mayrhofer, 2017; Meder, Mayrhofer, & Waldmann, 2014; Waldmann, 2000; Waldmann & Holyoak, 1992), legal and moral reasoning (Lagnado

[3]The other major approach to causality takes counterfactual conditionals as basic; see Lewis (2000).

& Gerstenberg, 2017; Wiegmann & Waldmann, 2014), conditional reasoning (Ali, Chater, & Oaksford, 2011; Fernbach & Erb, 2013; Hall, Ali, Chater, & Oaksford, 2016; Oaksford & Chater, 2017), and analogical reasoning (Holyoak & Lee, 2017; Holyoak, Lee, & Lu, 2010). Insofar as explanation is *causal* explanation, work on causal Bayes nets may also help to capture the notion of explanation, and to quantify explanation quality.[4]

To be sure, not every explanation is a causal explanation. Explanation in mathematics is perhaps the clearest example of non-causal explanation (Mancosu, 2015), but in the empirical sciences, too, there are types of explanation that are non-causal (Lange, 2017; Wouters, 1995, 2007). If Kitcher (1981) and others are right, then at least some, and possibly even all, types of non-causal explanation can be understood in terms of coherence (internal coherence, coherence with the data, coherence with background knowledge). Inasmuch as the currently best accounts of coherence are all of a probabilistic variety (e.g., Bovens & Hartmann, 2003; Douven & Meijs, 2007; Fitelson, 2003; Olsson, 2002), such types of explanation can still be analyzed in strictly probabilistic terms. In short, IBE as a rule of belief change different from Bayes' rule may well be perfectly compatible with Bayesianism as this position is understood in psychology.

All this is not to say that psychologists have been very actively researching versions of IBE. Indeed, the descriptive adequacy of IBE was, at least until recently, mostly unexplored territory. To some extent this is surprising, given that, as mentioned above, there is already a wealth of studies showing that explanation plays a number of distinctive roles in human cognition. But

[4]Lombrozo and Vasilyeva (2017, p. 418) suggest that explanatory considerations may actually inform causal inference, rather than the other way around; see, in a similar vein, Legare, Sobel, and Callanan (2017) and Pacer, Williams, Chen, Lombrozo, and Griffiths (2013). Douven and Schupbach's (2015b) findings suggest that it may pay off in this context to distinguish between objective and subjective probabilities (whenever objective probabilities are available). They were able to predict participants' subjective probability updates best by using objective probabilities together with measures of explanatory power *derived* strictly from those same objective probabilities. Connecting this result, and Lombrozo and Vasilyeva's suggestion, with research on causal Bayes nets goes beyond the scope of this paper, as does the question of the ordering of conceptual or psychological priority among the triad probability–causality–explanation.

none of those studies dealt with the question of the existence of a systematic, let alone rule-governed, relation between explanation and belief change.

As was also mentioned, Douven and Schupbach (2015a) did consider this question, specifically the question of how explanatory considerations affect probabilistic belief change (i.e., change of subjective probabilities). They report empirical work showing that their participants' assignments of subjective probabilities were influenced by such considerations indeed, even to the extent that taking into account their participants' explanatory judgments on some hypotheses in light of new evidence allowed Douven and Schupbach to predict with great accuracy how, on the basis of this evidence, the participants would change their probabilities for the hypotheses, and to predict this with greater accuracy than could be done purely on the basis of resources acceptable from a (strictly, in the philosophers' sense) Bayesian perspective.

Douven and Schupbach further reported that their participants' subjective probabilities were affected by their assessment of the *difference* in explanation quality between the two hypotheses whose probabilities they were asked to estimate. In particular, the better an explanation of the evidence one hypothesis was relative to the other hypothesis, in the perception of a participant, the greater the boost in that participant's confidence that the evidence tended to give to the former.

Because Douven and Schupbach's (2015a) work only concerned a probabilistic version of IBE, their results do not preempt any of our current research questions, which are about the relation between explanation and categorical (non-probabilistic) acceptance. Nevertheless, that those results showed explanatory considerations (and in the case of competing explanations, also difference in quality between those explanations) to play a role in people's reasoning is at least a preliminary reason for expecting positive answers to Q1 and Q3.

We found additional reason to expect positive answers even to all three questions in research that is *not* primarily concerned with the relation between explanation and inference. Tenney, Cleary, and Spellman (2009) are interested in the question of how to explicate the notion of being beyond reasonable doubt as it is used in criminal law, and specifically in whether the in-

troduction of additional suspects by the defense has an effect on people's judgments concerning the guilt (beyond reasonable doubt) of a target suspect. They presented their participants with a fictional murder case that had one main suspect and asked the participants to give a yes/no verdict concerning the guilt of this suspect; participants were also asked to indicate how likely it was, in their opinion, that the suspect had committed the murder. Tenney and colleagues varied between participants the number of alternative suspects, from 0 to 3 alternatives. They found that suggesting one alternative suspect significantly reduced the number of "guilty" verdicts and that it also somewhat lowered the participants' subjective probabilities for the main suspect's guilt, but that introducing any further suspects had little to no additional effect.[5]

While Tenney and colleagues make no explicit reference to the literature on explanatory reasoning, they do relate their findings to Pennington and Hastie's (1986, 1992) so-called story model; and as Byrne (1995) points out, although Pennington and Hastie are strictly concerned with legal contexts (specifically with how jurors arrive at a verdict), their model is otherwise close to Thagard's (1989) theory of explanatory coherence—which is an account of IBE of sorts. In Pennington and Hastie's experimental work, the emphasis is on processing, notably on how jurors try to construct complete and coherent stories out of the evidence laid out before them; how this story-building process is influenced by the order in which the various pieces of evidence are introduced; and on how jurors determine, in the end, how convincing the resulting story or stories are.

Most of Pennington and Hastie's work on the story model either preceded or more or less coincided with the publication of Lipton's and other philosophers' writings on IBE and specifically on contrasting explications of that rule of inference. Seen from a perspective informed by

[5]Dealing with the introduction of alternative theories at a more general level, Hemmerich, Van Voorhis, and Wiley (2016, Exp. 3) found that presenting participants with an alternative theory led to a significant drop in the confidence in the antecedently held theory when participants were also shown evidence undermining the latter, but not to a significant drop in acceptance of the antecedent theory. Given that Hemmerich and colleagues did not ask their participants to rate the explanation quality of the theories involved, their results are difficult to compare with those from our experiments.

the more recent philosophical theorizing on IBE, however, the outcomes of Tenney et al.'s study appear rather unsurprising, although there is prima facie more than one way to make good sense of them.

Even the staunchest advocates of IBE acknowledge that if we can infer to the only (reasonable) explanation of the available evidence, we are generally on safer ground than if we have to infer to the best explanation (see Bird, 2010). And in the Tenney et al. study, we go from an inference to what participants may well have thought of as the only plausible explanation, to a less secure inference to the best explanation when a second suspect besides the initial suspect is introduced. However, when still further suspects are then introduced, there is no further change in the rule of inference that one is relying on; that remains IBE. In other words, from the current perspective, one would have predicted that the introduction of a third or even a fourth candidate explanation of the crime—in the form of a third or fourth suspect—would have little to no additional effect on either people's willingness to accept, or their confidence in, the best explanation.

Another possibility is that, in the condition in which it was presented alongside a second explanation, the target explanation in the Tenney et al. study—the guilt of the prime suspect—appeared less good than in the condition in which it was presented alone. We have long known about contrast effects in perception—for instance, that one and the same color can look brighter in the presence of a second, less bright color than it looks on its own—but more recently these effects have been shown to occur also in the cognitive domain, and in particular to affect people's standards of judgment. For instance, Shoots-Reinhard, Rucker, Petty, and Shakarchi (2014) show that whether a consumer product is judged to be desirable may depend on which other products a potential buyer is attending to. A similar effect may have played a role in the Tenney et al. study, where it may or may not have been complementary to an effect brought about by the distinction between inference to the only versus to the best explanation.

As stated, Tenney et al.'s main concern was with the notion of being beyond reasonable doubt, and so it is understandable that they did not ask their participants for judgments of the *quality* of

the various explanations—featuring the various suspects—involved in their materials. Furthermore, Tenney et al. explicitly tried to make every suspect appear equally plausible as the murderer, mentioning varying the materials on this count as an avenue for future research (which, to the best of our knowledge, they have not further explored to this date). Varying the plausibility of the suspects might in fact have given a clear indication of whether their results are indeed due to how people's reasoning is guided by explanatory considerations, at least supposing Bird's previously mentioned suggestion to be correct. After all, that suggestion implies that the drop in percentage of "guilty" verdicts that is present in Tenney et al.'s data would have been smaller, or even altogether absent, if the second-most-plausible suspect had been markedly less plausible than the target suspect.

In the first two experiments now to be reported, we presented people with events and accompanying potential explanations of those events and then asked one group of participants for their judgments of how good the explanations are,[6] and another group whether they agreed with the explanations. We followed Tenney and colleagues in asking the second group also how probable they deemed the explanations, as this would enable us to compare explanation quality and probability as predictors of agreement, and hence to investigate a sharpening of question Q1: Is the quality of an explanation a good predictor of people's readiness to accept it, and is it a better predictor than the explanation's probability?

[6] Vasilyeva, Wilkenfeld, and Lombrozo (2017) found that judgments of explanation quality can depend on what task an evaluator is faced with in a given context. For instance, an evaluator may judge a mechanistic explanation of a phenomenon to be better than a functional explanation when she is faced with a task that makes knowledge of underlying mechanisms more important than knowledge of functional connections. It is reasonable to suppose that this kind of sensitivity to contextual relevance has played no role in our studies, given that in none of the studies were participants presented with anything like the tasks given to the participants in Vasilyeva et al.'s studies.

Experiment 1

Method

Participants

Three hundred and twenty-one participants were recruited on CrowdFlower, where they were directed to the Qualtrics platform, via which the survey was administered. They were financially compensated for the time and effort spent on the survey. We excluded from analysis non-native speakers of English (given that all materials were in English) as well as participants who failed either of two validation checks. The first check, which appeared at the end of the demographic section, was a question taken from Pennycook, Trippas, Handley, and Thompson (2014). This question showed a list of hobbies and asked, "Below is a list of hobbies. If you are reading these instructions please write 'I read the instructions' in the 'other' box." We excluded data from participants who failed to enter the requested words. The second validation question was taken from Aust, Diedenhofen, Ullrich, and Musch (2014) and appeared at the end of the study. Participants were asked if they had responded seriously to the questions in the experiment, emphasizing that their answer would not affect payment. Those who responded in the negative were excluded. Finally, we removed participants who completed the survey in less than 2 minutes, which pilot testing had shown to be the minimally required time to read all the materials and questions. This left us with 275 participants for the final analysis. These remaining participants spent on average 387 seconds (SD = 468) on the survey; time spent on survey was non-normally distributed, with skewness of 7.69 (SE = 0.15) and kurtosis of 68.76 (SE = 0.30). One hundred and fifty-eight of the remaining participants were female, 115 male, and 2 preferred not to respond to the question concerning gender. Their mean age was 38 (SD = 12). Finally, 213 of them had a college degree, 60 indicated high school as their highest education level, and 2 indicated a lower education level. Repeating the analysis using the data from all of the initial 321 participants yielded qualitatively identical results.

Materials

Our materials were modeled on Tenney et al.'s murder case. In all three experiments, we used the same six basic scenarios, the first of which is a shortened version of the murder case used in Tenney et al.'s experiment, and the other five of which are structurally very similar. As we saw, Tenney and colleagues varied the number of alternative explanations offered to their participants. Our first experiment is concerned with the case of zero alternatives, meaning that the materials offered only one explanation in each scenario.

In this experiment, each of our six basic scenarios appeared in three different versions. Here is one version of what is essentially Tenney et al.'s murder case:

> Mrs. Smith, a high-ranking administrator from a top-tier university, was found strangled in her office. She had been in the process of divorcing her husband, Mr. Smith, because she had fallen in love with another man and wanted to pursue this new relationship. Both she and her husband were seeking the custody of their two children. A surveillance video showed Mr. Smith leaving the building in which his wife's office is located approximately 30 minutes before Mrs. Smith's body was discovered. Mr. Smith had in the past also been accused of domestic violence, in particular connected with his very strong jealousy.

One group of participants who were shown this version were then asked, "Do you agree that it was Mr. Smith who strangled Mrs. Smith?"; they could answer either "yes" or "no" by clicking a radio button. After they had answered this question, these participants were further asked to indicate on a scale from 0 to 100 percent how likely they thought it was that Mr. Smith had strangled Mrs. Smith, where they had to enter their answer in a text box. In the general instructions, they were informed that, on the given scale, "100 means that you are 100 percent certain that the event happened and 0 that you are 0 percent certain that it happened (so 100 percent certain that it did not happen)." The second group of participants who saw this version were asked to rate, on a 7-point Likert scale with the anchors labeled "Very bad" and "Very good" and the midpoint labeled "Neither good nor bad," the quality of the possible explanation that Mr. Smith murdered Mrs. Smith.

In the above example, the suggested explanation of Mrs. Smith's murder is—we supposed—relatively strong. The second version of the same scenario suggested what in our pre-theoretical judgment was a somewhat weaker explanation instead:

> Mrs. Smith, a high-ranking administrator from a top-tier university, was found strangled in her office. Mr. Hanson, one of Mrs. Smith's coworkers, had a crush on her, but she had turned him down. In fact, she had done so in a rather rude way. Mr. Hanson had been depressed since, and had started drinking heavily. In the past few weeks, Mrs. Smith had complained to Human Resources twice about Mr. Hanson insulting her and threatening her during work hours.

Again, this was shown to two groups, where these were asked the same questions concerning agreement-and-likelihood and explanation quality, respectively, that were presented to the groups who saw the stronger explanation.

Two further groups were shown a still weaker explanation, and were again asked either whether they agreed with the explanation and how likely they thought it was, or, to rate the quality of the explanation. Here is the version of the Mrs. Smith scenario shown to these groups:

> Mrs. Smith, a high-ranking administrator from a top-tier university, was found strangled in her office. One possibility is that she stumbled on the carpet as she was putting her scarf on and accidentally strangled herself to death. The scarf that she normally wears was found in her handbag and not around her neck.

Two other scenarios were also crime stories, one about the murder of Lady Windermere, and the other about a stolen Rembrandt painting; like the above scenario, these other crime stories feature mainly intentional relations, and mechanistic relations only to a lesser extent. Then there was a story about a broken dam, one about a malfunctioning camera, and one involving a medical diagnosis, with a suspicion of tuberculosis. These last three scenarios mainly involve causal and functional relations. See the Appendix for the full materials.

Design and procedure

We used a 3 (explanation-quality type: strong / intermediate / weak) × 2 (question type: agreement-and-likelihood / explanation quality) design, with conditions being manipulated between participants. Participants were randomly assigned to one of the six resulting groups (N_{SA} = 49, N_{SE} = 47, N_{IA} = 38, N_{IE} = 48, N_{WA} = 47, and N_{WE} = 46, with S, I, and W denoting the strong, intermediate, and weak condition, respectively, and with A denoting the agreement-and-likelihood condition and E the explanation-quality condition). Each participant was shown all six scenarios (in the version corresponding to the participant's explanation-quality condition), and each scenario was shown on a separate page (i.e., screen), in an order randomized per participant.

Results

The analysis consisted of two parts. In the first, a mixed-effects models approach was used to investigate the effect of explanation-quality type (hereafter, type) on the individual responses. In the second part, ordinary least squares models were used to analyze relationships between responses per scenario, aggregated across participants.

Mixed-effects models. In this part, we fitted three mixed-effects models, all with type as fixed effect, and with crossed random effects for participants and scenarios and with by scenarios random slopes for type. Specifically, we fitted two linear mixed models (LMM; see Baayen, Davidson, & Bates, 2008, or Pinheiro & Bates, 2000), one with explanation-quality ratings as dependent variable and the other with probabilities as dependent variable, and we fitted a binomial generalized linear mixed model (GLMM; see Jaeger, 2008, or Stroup, 2012) with the yes/no responses to the agreement question as dependent variable. The models were fit using the `lme4` package (Bates, Mächler, Bolker, & Walker, 2015) for the statistical computing language R (R Core Group, 2017); significance tests for the fixed effects were obtained via the `afex` package (Singmann, Bolker, Westfall, & Aust, 2017). Figure 1 plots the data together with a graphical summary of the main results from the analysis.

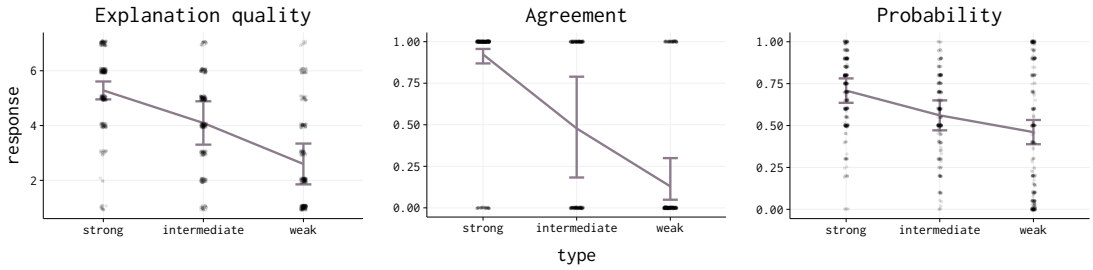

Figure 1: Estimated marginal means obtained from the mixed-effects models from Experiment 1, with error bars indicating 95-percent-confidence intervals; responses are plotted with jitter added to enhance visibility.

The first LMM showed that there was a significant effect of type on explanation-quality ratings, $F(2, 7.73) = 18.56$, $p = .001$, with follow-up tests using the `lsmeans` package (Lenth, 2017) and assuming Bonferroni correction showing that in the weak condition explanation quality was rated significantly lower (estimated marginal mean, or EMM, of 2.60) than in the strong condition (EMM = 5.28), $p = .0004$, and also than in the intermediate condition (EMM = 4.09), $p = .03$; the difference between the EMMs in the strong and intermediate conditions was borderline significant, $p = .054$. One-way ANOVAs conducted per scenario showed that mean ratings in the strong condition were always higher, and almost always significantly higher, than mean ratings in the intermediate condition, and were always significantly higher than mean ratings in the weak condition; for all scenarios, mean ratings in the intermediate condition were significantly higher than mean ratings in the weak condition.

Furthermore, there was a significant effect of type on probabilities, $F(2, 51.65) = 12.00$, $p < .0001$, follow-up tests showing that the explanations in the strong condition were deemed significantly more probable (EMM = 0.71) than those in the intermediate condition (EMM = 0.56), $p = .03$, and those in the weak condition (EMM = 0.46), $p < .0001$. The difference between the intermediate and weak conditions was not significant, $p = .15$. A series of six one-way ANOVAs showed that this pattern basically replicated per scenario.[7]

Finally, the GLMM revealed a significant effect of type on participants' willingness to agree with an explanation, $\chi^2(2) = 14.13$, $p < .001$, pairwise comparisons showing that the probabil-

[7]The function `ptt` in the online R script can be used to obtain details per scenario.

Table 1: Comparison of the regression models from Experiment 1.

	k	LL	AIC	BIC	R^2
ME	3	−61.49	128.97	131.65	.95
MP	3	−68.76	143.51	146.18	.88

Note: k is the number of parameters and LL the log-likelihood. AIC is the Akaike Information Criterion and BIC the Bayesian Information Criterion, two metrics that weigh model fit against model complexity. Their values are to be used comparatively, in that models with smaller values are taken to be predictively more accurate than ones with larger values. R^2 is the correlation between fitted and observed values.

ity of agreement in the strong condition (EMM = .92) differed significantly from the probability of agreement in the intermediate (EMM = .48) and weak (EMM = .13) conditions, $p = .0019$ and $p < .0001$, respectively; the difference between the probabilities of agreement in the intermediate and weak conditions was also significant, $p = .04$. Ordinary generalized linear models conducted per scenario found the same pattern in all six cases.

Ordinary least squares models. In the second part of the analysis, we asked whether mean judgments of explanation quality, which were based on the responses from one group of participants, had any predictive value for the percentages agreement obtained from the other group of participants. We also wanted to know how the predictive value of mean judgments of explanation quality from the one group of participants compared to the predictive value of the mean probabilities for the explanations from the other group. To answer these questions, we fitted two linear models, both with percentage agreement (18 data points, from 6 scenarios judged in 3 conditions) as dependent variable, and one of them (ME) with mean judgments of explanation quality as independent variable and the other (MP) with mean probabilities as independent variable. Table 1 gives the most important outcomes of the model comparison, and Table 2 gives the regression results. We see that while, in terms of R^2-value, both models do quite well, ME outperforms MP across all criteria. According to Burnham and Anderson (2002, p. 70 f), a difference in AIC-value greater than 10—as is the case here—is to be interpreted as indicating that the model with the higher value receives no support from the data.

It appears that judgments of explanation quality as given by one group of participants is a more reliable predictor of the extent to which a second group of participants agrees with the

Table 2: Regression results for ME and MP.

		b	SE	*β*	*t*	*p*
ME	Intercept	−42.78	6.04		−7.09	< .0001
	E	23.99	1.44	0.97	16.67	< .0001
MP	Intercept	−97.94	14.34		−6.83	< .0001
	P	2.61	0.24	0.94	10.72	< .0001

explanations than the probabilities assigned to those explanations by that group. To repeat a point made earlier, we do not believe that the issue of the predictive value of explanatory judgments bears directly on Bayesianism as this position is commonly understood among psychologists; but for hard-nosed Bayesian philosophers the aforementioned finding might be cause for concern. On the other hand, these Bayesians might feel entitled to reject the result, given that they reject the statistics used in the analysis. We therefore also conducted a Bayesian regression analysis, basically refitting the models ME and MP using the `BayesFactor` package (Morey & Rouder, 2015), which yielded a Bayes factor of 2.5×10^8 for the model with mean explanatory judgments as predictor and a Bayes factor of 6.5×10^5 for the model with mean probabilities as predictor. A Bayes factor quantifies the degree to which the data favor a given model over the null (intercept-only) model. In the present case, the Bayes factors indicate that the data very strongly support both models compared to the null model (Jeffreys, 1961, p. 432; also Kass & Raftery, 1995). More importantly, they show that the data favor the model with explanatory judgments over the model with probabilities by a factor of approximately 385.[8,9]

[8] Where M_0 is the null model and A the percentages agreement from our experiment, to say that the Bayes factor for ME equals 2.5×10^8 is to say that $\Pr(A \mid ME)/\Pr(A \mid M_0) = 2.5 \times 10^8$. The extent to which A favors ME over MP is given by $\Pr(A \mid ME)/\Pr(A \mid MP) = \big(\Pr(A \mid ME)/\Pr(A \mid M_0)\big)/\big(\Pr(A \mid MP)/\Pr(A \mid M_0)\big) = 384.62$.

[9] A referee raised the concern that, by asking the participants in the agreement-and-likelihood condition whether they *agreed* with the explanation, rather than whether they *believed* or *accepted* it, we may have suggested too strongly that we ourselves, or in any case someone, already believed the explanation to be true. The same referee wondered whether participants might not have taken the phrase "one possibility is that," as used in the weak version of the Mrs. Smith scenario, as a marker of moderate or even low confidence. To address these concerns, we ran a control experiment using the weak versions of all scenarios, now asking whether the participants believed the explanations, and replacing the potentially problematic phrase, "One possibility is that she stumbled on the carpet . . . ," by

Discussion

At design stage, we had classified the various explanations into three groups, according to how satisfactory they appeared to us. The mixed-models analysis of the explanation-quality ratings showed participants' judgments to be in broad agreement with our own intuitions of explanation quality. The other two mixed models showed that explanation-quality type affected not only explanation-quality ratings but also probabilities and agreement.

The second part of the analysis showed that while probabilities were an excellent predictor of agreement in one group of participants, the explanation-quality ratings of another group of participants predicted agreement in the first group still more accurately. This result is in line with Douven and Schupbach's (2015a) finding that objective probabilities predicted accurately their participants' change in subjective probabilities, but that explanation-quality judgments were a better predictor still.

We also found some support for the descriptive adequacy of Lipton's idea that an explanation must be good enough to be acceptable. Given that our participants were always offered a single explanation, they might conveniently have gone with that one, perhaps also judging it to be good, simply because of lack of contrast. But that is not what we found. In the weak condition, most participants who were asked to judge the quality of the explanations judged them to be precisely that: weak. Correspondingly, most of the participants who were asked whether they agreed with these explanations answered negatively.

"Mrs. Smith could have stumbled on the carpet . . . " To analyze the results, we fitted a GLMM with the responses to the agreement question in the weak condition from Experiment 1 and the responses to the belief question from the control experiment as dependent variable, with question type (agree / believe) as fixed effect, and with crossed random effects for participants and scenarios with by scenarios random slopes for question type. The effect of question was not significant, $\chi^2(1) = 0.42$, $p = .52$. We further fitted an LMM, this with the responses to the corresponding probability questions as dependent variable, and here too, question type was not significant, $F(1, 81.69) = 0.17$, $p = .68$. Fitting parallel Bayesian models revealed substantial support for the null hypothesis, in both cases. See the Supplementary Information for more details.

On the other hand, Fernbach, Darlow, and Sloman (2010) found that while people tend to ignore possible alternative explanations in *predictive* reasoning—when they reason from an explanation to a possible effect—they are more likely to conceive of alternatives in *diagnostic* reasoning, that is, when they reason from an effect to a possible explanation. Indeed, it makes sense to think that, the poorer the only given explanation is, the likelier people are to engage in the process of looking for alternative explanations. As a result, especially the participants in the weak condition may well have considered other explanations, and possibly better ones, than the one they were presented with. This means that the present data at best weakly support Lipton's idea. We relegate further discussion of this idea to Experiment 3.

Experiment 2A

We now turn to the question of the effect of competing potential explanations. If there is a best and good-enough explanation for the evidence at hand, then it is still an open question how much better, qua explanation, it is than its closest competitor. As Bird (2000) suggested, and as was also suggested by the results from Douven and Schupbach (2015a), this question may matter to what we should infer from the evidence.

The questions to be addressed are the following: Can we, using our materials, replicate Tenney, Cleary, and Spellman's (2009) finding of a second suspect leading to a significant decrease in "guilty" verdicts? Supposing we can, why does the effect occur? Is it because when contrasted with a second explanation, the target explanation comes to appear less good? Or is the target explanation perceived as about equally good (or equally bad) as it is when presented alone, but participants become more hesitant in accepting it because the inference they envisage making is no longer an inference to the only (reasonable) explanation but an inference to the best explanation? The latter would probably be felt to be a less secure type of inference than the former.

Method

Participants

Two hundred fourteen participants were recruited in the same way as in Experiment 1. They were financially compensated for their time and effort. We used the same validation checks and exclusion criteria as in the previous experiment, which left 187 participants for the final analysis. These participants spent on average 510 seconds (SD = 406) on the survey; time spent on the survey was again non-normally distributed, with skewness of 7.04 (SE = 0.18) and kurtosis of 70.46 (SE = 0.36). Of the remaining participants, 101 were female, 85 male, and 1 chose not to answer the question about gender. Their mean age was 37 (SD = 13). One hundred and thirty-four participants had a college education, 46 indicated high school as their highest education level, and 7 indicated a lower education level. An analysis conducted without excluding any participants led to very similar results.

Materials

The materials consisted of the same six basic scenarios that were used in Experiment 1, but now the participants were shown versions with *two* explanations instead of one. The first of these explanations—which we refer to as "the target explanation"—was the same as the explanation in the corresponding scenario in the strong condition from the previous experiment. Then one group was shown an additional explanation that we, the experimenters, believed to be roughly as plausible as the target explanation, and the other group was shown as second explanation the one for the corresponding scenario in the intermediate condition of Experiment 1; hence, a somewhat less plausible explanation than the target explanation (not only in the experimenters' judgments, but also as confirmed by the results from Experiment 1). For instance, here is the version of what is essentially Tenney et al.'s murder case with two strong explanations:

> Mrs. Smith, a high-ranking administrator from a top-tier university, was found strangled in her office.

She had been in the process of divorcing her husband, Mr. Smith, because she had fallen in love with another man and wanted to pursue this new relationship. Both she and her husband had been seeking the custody of their two children. A surveillance video showed Mr. Smith leaving the building in which his wife's office is located approximately 30 minutes before Mrs. Smith's body was discovered. Mr. Smith had in the past also been accused of domestic violence, in particular connected with his very strong jealousy.

However, Mrs. Smith also knew about extremely incriminating evidence against Mr. Hanson, one of her coworkers. The evidence is in fact so damaging that it could lead to the termination of Mr. Hanson's contract. Mr. Hanson was desperate to keep Mrs. Smith from revealing his secret. On the day of the murder, Mr. Hanson was in his office, which is close to Mrs. Smith's office.

In the corresponding version with the less plausible alternative explanation, the paragraph about Mr. Hanson was replaced by the text cited in the materials section of the previous experiment, specifically by the part starting "Mr. Hanson, one of Mrs. Smith's coworkers ..." and then down to the end, exactly as stated in the previous section. By normal standards, in the first case Mr. Hanson had more of a motive for murdering Mrs. Smith than in the second.

Each of the two groups was again split into two, with one subgroup of each group being asked whether they agreed that Mr. Smith had strangled Mrs. Smith and also how likely they thought that was, and the other subgroup being asked to rate the quality of both possible explanations—that Mr. Smith murdered Mrs. Smith, and that Mr. Hanson did so. Again, the Supplementary Information contains all materials used in this experiment.

Design and procedure

We used a 2 (version of scenario-pair: strong target–strong alternative / strong target–intermediate alternative, or for brevity, strong / intermediate) × 2 (question type: agreement-and-likelihood / explanation quality) design. The conditions were manipulated between participants, resulting in four groups ($N_{SA} = 43$, $N_{SE} = 44$, $N_{IA} = 52$, and $N_{IE} = 48$, with S denoting the strong

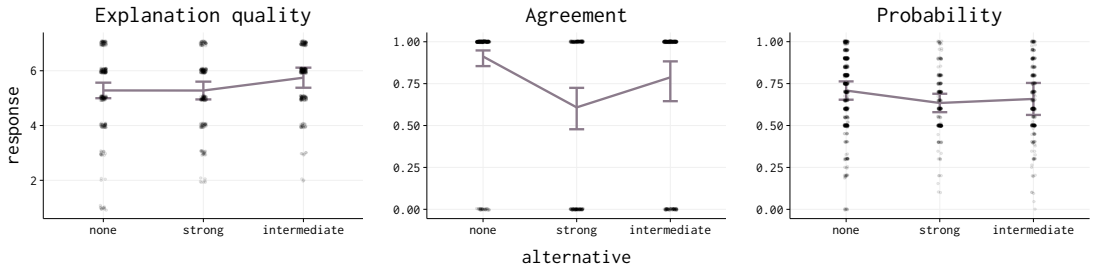

Figure 2: Estimated marginal means for target explanations, obtained in the mixed-models analysis of the data from Experiment 2A (yielding the EMMs for the strong and intermediate conditions) merged with the data from the strong condition from Experiment 1 (yielding the EMM for what is the none condition here); error bars indicate 95-percent-confidence intervals, and responses are plotted with added jitter to enhance visibility.

condition, etc.). Each scenario was shown on a separate page in an order randomized per participant.

Results

Because we were not only interested in comparing the strong and intermediate conditions with each other but also with the strong condition from Experiment 1, in which the same target explanation had been presented, but unaccompanied by an alternative, we merged the responses (the explanation-quality ratings as well as the responses from the agreement-and-likelihood group) from the strong condition of the previous experiment with the newly collected responses. The analysis then consisted of two parts, mirroring the analysis of the data obtained in Experiment 1.

Mixed-effects models. We started again by fitting three mixed-effects models, two LMMs with the explanation-quality ratings and the probabilities, respectively, as dependent variable, and a binomial GLMM with the responses to the agreement question as dependent variable; all three models had alternative (levels: none / strong / intermediate) as fixed effect factor, and included crossed random effects for participants and items as well as by items random slopes for alternative. Figure 2 presents a graphical summary of the main findings, plotted together with the merged data.

The first model revealed that there was no significant effect of alternative on explanation-quality ratings, $F(2, 51.11) = 3.00$, $p = .06$, meaning that there is no evidence that people rate

the quality of an explanation differently depending on whether the explanation is presented on its own or together with an alternative, be it a strong one or one of intermediate strength. The second model showed that neither was there a significant effect of alternative on probabilities, $F(2, 13.58) = 3.49$, $p = .06$. By contrast, the GLMM showed that the effect of alternative on agreement was highly significant, $\chi^2(2) = 16.90$, $p = .0002$. Pairwise comparisons revealed that the probability of agreeing with an explanation in the none condition (EMM = .91) was significantly higher than the probability of agreeing in the strong condition (EMM = .61), $p < .0001$, and also than the probability of agreeing in the intermediate condition (EMM = .79), $p = .012$; the probability of agreeing in the intermediate condition was in turn significantly higher than the corresponding probability in the strong condition, $p = .024$. Further follow-up analyses, with generalized linear models conducted per scenario, revealed the same pattern for all six scenarios.

Ordinary least squares models. In the first experiment, we asked how reliably acceptance rates can be predicted on the basis of perceived explanation quality. For the data from Experiment 2A, we can ask that question again. However, now we can also look into the predictive value of perceived difference in quality between target and alternative explanation. Specifically, we can ask whether people's agreement (or otherwise) with a given explanation depends, if only partly, on the quality of another explanation that they are presented with. So, we now fitted four linear models, all with

Table 3: Regression results for MTA, MT, MD, and MP.

		b	SE	β	*t*	*p*
MTA	Intercept	−35.23	83.95		−0.39	.70
	T	21.66	12.93	0.59	1.68	.13
	A	−4.09	4.85	−0.30	−0.84	.42
MT	Intercept	−105.38	33.01		−3.19	.01
	T	31.28	6.01	0.85	5.20	.0004
MD	Intercept	55.48	3.02		18.40	< .0001
	D	8.68	1.71	0.85	5.09	.0005
MP	Intercept	−33.15	21.21		−1.56	.15
	P	1.52	0.32	0.83	4.70	.0008

Table 4: Comparison of the ordinary least squares models from Experiment 2A.

	k	LL	AIC	ΔAIC	BIC	ΔBIC	R^2	BF
MTA	4	−39.72	87.45	1.09	89.38	1.57	.75	21.28
MT	3	−40.18	86.36	0.00	87.81	0.00	.73	57.59
MP	3	−41.05	88.09	1.73	89.55	1.74	.69	32.12
MD	3	−40.37	86.74	0.38	88.19	0.38	.72	50.61

Note: For explanation, see the note to Table 1. NB: The BF value for a given model pertains in actuality to the Bayesian model with the same dependent and independent variables.

percentage agreement as dependent variable, and one (MT) with mean judgments of quality of the target explanation (T) as independent variable, one (MTA) with mean judgments of quality of both the target explanation and the alternative explanation (A) as independent variables, one (MD) with the difference between those means (D) as independent variable, and one (MP) with mean probabilities (P) as independent variable. The results from these regressions are stated in Table 3. For the reasons explained previously, we conducted Bayesian regressions in parallel.

Results of the model comparisons are stated in Table 4. They are in agreement with what we found earlier to the extent that, across all criteria, the probabilistic model does worst. At the same time, the differences between the four models are rather small. Judging by AIC and BIC as well as by the Bayes factors, MT does best, closely followed by MD. Looking specifically at Bayes factors, we can say that the data only very slightly favor MT over MD (by a factor of 1.1), while either of these models receives close to twice as much support from the data as MP does. So, just as we found in Experiment 1, mean probabilities are seen to be a worse predictor of agreement than judgments of explanation quality, even though, like previously, the probabilities came from the same group of participants who were asked whether they agreed with an explanation while the judgments of explanation quality came from a different group of participants.

Finally, in the first part of the present analysis we saw that there was a drop in agreement when the target explanation was presented alongside a strong alternative explanation as compared with when it was presented alone, in the first experiment, and that there was also a drop, though a smaller one, when the target explanation was presented alongside a somewhat less strong alternative. Given our data about the differences in perceived quality between a target

Table 5: Regression results for MDE and MDP.

		b	SE	*β*	*t*	*p*
MDE	Intercept	30.29	2.01		15.06	< .0001
	DE	−7.88	1.14	−0.91	−6.93	< .0001
MDP	Intercept	12.63	3.49		3.62	.005
	DP	1.40	0.42	0.72	3.31	.008

explanation and its strong alternative as well as between the same target and its less strong alternative, can we predict the drop in agreement with that target that was brought about by going from an unaccompanied presentation (in Experiment 1) to an accompanied presentation (in Experiment 2A)? And how does the predictive accuracy of difference in explanation quality compare with the predictive accuracy of difference in assigned probabilities?

To answer these questions, we fitted the following linear models: (i) a model (MDE) with, as dependent variable, difference in percentage agreement with the target explanation when it was presented alone and when it was presented with an alternative, and, as independent variable, difference in explanation quality between target and alternative (DE), and (ii) a model (MDP) with the same dependent variable but with difference in assigned probability when presented alone—as recorded in Experiment 1— and when presented with an alternative (DP) as independent variable. The relevant information about these models is given in Tables 5 and 6; the latter also states the Bayes factors from the parallel Bayesian regressions we conducted. Most notably, the estimated coefficient for DE in MDE reveals that for every Likert-scale point that the difference in quality between target and alternative explanations becomes *larger*, the drop in agreement rates becomes, on average, smaller by about 8 percent (so there is a *decrease* in the *decrease*: as the difference in quality increases, there is still a drop, but it gets less big).

To summarize the information given in the tables, we can say that difference in quality be-

Table 6: Comparison of MDE and MPD.

	k	LL	AIC	BIC	R^2	BF
MDE	3	−36.65	79.31	80.76	.81	372.02
MDP	3	−42.43	90.85	92.31	.51	6.13

Note: For explanation, see the note to Table 1. NB: The BF values pertain to the corresponding Bayesian models.

tween target and alternative explanation allows us to accurately predict the difference in agreement with the target explanation as shown alone in Experiment 1 and as shown with the alternative in Experiment 2A, and to predict this much more accurately than we could do on the basis of difference in probabilities. In other words, knowing the difference in quality between two explanations, we can accurately predict what the drop in acceptance will be when the better of those explanations is presented in tandem with the other, as opposed to when it is presented alone. In this regard, the predictive accuracy of difference in explanation quality is much greater than the predictive accuracy of difference in probability.

Discussion

We found confirmation for the descriptive adequacy of what Bird meant as a normative proposal, viz., that the difference in quality between the best explanation and its closest competitor matters to the acceptability of the former. In particular, we saw that presenting the strong explanations from Experiment 1 alongside an alternative explanation led to a significant drop in agreement when that second explanation was itself relatively strong, whereas the drop was much smaller when the second explanation was weaker. It was also seen that difference in explanation quality accurately predicts difference in agreement with the target explanation as shown alone in Experiment 1 and as shown with an alternative in Experiment 2A, and predicts this much more accurately than difference in subjective probabilities does.

The results are not entirely surprising, in that they are in line with Douven and Schupbach's (2015a) findings; these authors reported that difference in explanation quality was an excellent predictor of people's probabilistic belief changes. Also, the drop in acceptance rates brought about by offering a second explanation was expected on the basis of Tenney et al.'s finding that introducing a second suspect led to a significant decrease of "guilty" verdicts. We asked earlier why that drop occurred in the Tenney et al. study and speculated that it might be (i) because we go from an inference to the *only* explanation to an inference to the *best* explanation of several, or

(ii) because the target explanation comes to look *less compelling* when presented together with a competitor.

Our results allow us to reject (ii) right away: the effects we found cannot be generally attributed to a difference in how people perceive the target explanation, depending on whether it appears alone or in the company of an alternative, because the mixed-effects model with explanation-quality ratings as dependent variable showed that there was no significant difference in perceived quality among the target explanations. As for (i), our results suggest that this is at best part of the explanation. After all, the mixed-effects model with responses to the agreement question as dependent variable showed that there was a significant drop in acceptance also in the intermediate condition, but that this drop was significantly smaller than the drop in acceptance in the strong condition. The difference cannot be explained by assuming that participants in the intermediate condition were really inferring to the only (reasonable) explanation, because we know from the first experiment that, even though the explanations that figured as alternatives in that condition are generally perceived as being not quite as strong as the target explanations, they were deemed far from implausible by most participants. Thus, why did the introduction of alternative explanations significantly reduce agreement, where the extent of reduction depended on the quality of the alternative?

Here is a suggestion that we aimed to investigate as part of Experiment 3. Williamson (2007, p. 224) notes that "[t]he concept of a better explanation is an informal one, rooted in ordinary ways of thinking, even if scientists' particular applications of it are informed by their background knowledge." It has often been said that explanation quality depends on such factors as simplicity, scope, and coherence with background knowledge (e.g., McMullin, 1996), but each of those factors carries some vagueness with it.[10] And explanation quality being not fully formal and somewhat vague, we may feel that our verdicts concerning the relative quality of explanations

[10]See Zemla, Sloman, Bechlivanidis, and Lagnado (2017) for a long list of possible attributes of explanations, and for experimental results concerning which of those attributes matter most to people in assessing the quality of everyday explanations.

are not entirely reliable whenever explanations are similar in quality. This is not completely speculative. Horry and Brewer (2016) report that when participants were briefly shown a face and then asked to identify that face in a set of faces shown simultaneously, the confidence in their choice correlated negatively with the degree of similarity between the target face and whichever other face in the set was most similar to it. Our suggestion is that much the same holds for explanations, where the relevant similarity then concerns the quality of the explanations: given two explanations competing for bestness, the more similar in quality they are, the less confident people become in their judgment of which is best, and the more hesitant they become to infer the truth of that explanation.

Experiment 2B

We saw that percentages agreement were better predicted by explanatory judgments than by probabilities, even though percentages agreement and probabilities were from the same group of participants while explanatory judgments were from a different group. Yet the comparison might be deemed not entirely fair, given that we recorded explanatory judgments for both the target and the alternative explanation but probabilities only for the target explanation. Might we not be able to predict percentages agreement still more accurately via probabilities if we had, in a completely symmetric fashion, also asked for participants' probabilities for the alternative explanations?[11] To answer this question, we conducted a follow-up study in which we did precisely that, then fitted a number of additional ordinary least squares models using the newly measured variables, and finally compared those models to the ordinary least squares models from the previous section.

[11] Thanks to Mike Oaksford for pressing us on this.

Method

Participants

One hundred and two participants were recruited in the same way as in the previous experiments. They received a small payment for their participation. We used the same validation criteria as in the previous experiments. We also applied the same exclusion criteria, which left us with 87 participants. In this experiment, participants were asked, per scenario, for two probabilities whose sum could not exceed 100 (though the probabilities could add up to less than 100). Participants were informed about this at the start of the survey and were explicitly reminded of it on every page. Further excluding participants whose probabilities summed to more than 100 left us with 77 participants for the final analysis. These participants spent on average 727 seconds on the survey (SD = 1783); time spent on survey was non-normally distributed, with skewness 8.40 (SE = 0.28) and kurtosis 72.58 (SE = 0.56). Forty-one of these participants were female, 36 were male. Their mean age was 39 (SD = 12), and 58 indicated college as their education level, 16 high school, and 3 a lower education level.

Materials, design, and procedure

Everything was exactly as in Experiment 2A limited to the explanation-quality groups, the only difference being that participants were now asked how likely, in their opinion, were the two explanations per scenario that were presented to them, and not to rate the quality of those explanations. So, in particular, there were only two groups in this experiment, one being shown the strong alternative explanations (the strong condition, N = 38) and the other being shown the somewhat less strong alternatives (the intermediate condition, N = 39).

Results and discussion

We fitted three ordinary least squares models, all with percentage acceptance as based on the responses from Experiment 2A as dependent variable, and one model (MPTA) with both probabilities for the target explanations and probabilities for the alternatives as predictors, one (MPT)

Table 7: Comparison of the regression models from Experiments 2A and 2B.

	k	LL	AIC	ΔAIC	BIC	ΔBIC	R^2	BF
MPTA	4	−41.11	90.23	3.87	92.17	4.36	.68	9.65
MPT	3	−41.17	88.34	1.98	89.79	1.98	.68	29.60
MPD	3	−41.47	88.94	2.58	90.40	2.59	.67	24.21
MTA	4	−39.72	87.45	1.09	89.38	1.57	.75	21.28
MT	3	−40.18	86.36	0.00	87.81	0.00	.73	57.59
MP	3	−41.05	88.09	1.73	89.55	1.74	.69	32.12
MD	3	−40.37	86.74	0.38	88.19	0.38	.72	50.61

Note: For explanation, see the note to Table 1. NB: The BF values pertain to the corresponding Bayesian models.

with only probabilities for the target explanations as predictor, and one (MPD) with difference between probabilities for targets and for alternatives as predictor.

Table 7 gives the relevant model comparison statistics. To facilitate comparison with the corresponding models from Experiment 2A, the statistics for those models are stated again here. We see that eliciting probabilities also for alternatives did not lead to any improvement in predictive accuracy. In particular, the new models do still worse, across all criteria, than the best models from the previous experiments, which predict agreement on the basis of explanation-quality judgments.

We could now also compare the model MDE from the analysis of Experiment 2A—which was the model that predicted drop in agreement when an alternative explanation was added on the basis of the difference in explanation quality between target and alternative explanation—with a similar model that has difference in probability between target and alternative as a predictor. In this comparison, MDE was a again far superior, the new model having an AIC value of 87.21 (vs. 79.31 for MDE), a BIC value of 88.67 (vs. 80.76), and R^2-value of .64 (vs. .81), and a Bayes factor of 17.41 (vs. 372.02).

Experiment 3

So far, we manipulated judgments of explanation quality and the question of agreement between groups. Therefore, all analyses aiming to shed light on how people's judgments of explanation

quality are related to their willingness to infer to the truth of the given explanation were conducted on mean responses. To obtain a deeper insight into this question, one will have to elicit judgments of explanation quality and agreement responses from the same participants. This was the main motivation for the third experiment. A subsidiary motivation came up in the analysis of the results from Experiment 2A, which suggested that the drop in percentage agreement when an explanation is presented in the company of another explanation was caused by a drop in people's confidence in their judgment of explanatory bestness, which can be regarded as a metacognitive issue in the manner of Ackerman and Thompson (2015, 2017a, 2017b). To see whether there is any truth to that suggestion, we made use of a tool, developed in the metacognition literature (Thompson, Prowse Turner, & Pennycook, 2011), for measuring people's metacognitive confidence.

Method

Participants

Eighty-four participants were recruited on Prolific Academic, where they were directed to the Qualtrics platform, on which the survey was run. Participants received a small amount of money in return for their cooperation. This time, because of the greater length of the experiment, there were three attention checks, which also served as distraction material between blocks of questions (see below). After removing participants who missed at least one of these checks, who returned incomplete response sets, or who indicated that they were non-native speakers of English, 70 participants remained. Because in this experiment participants had to read all scenarios twice, it seemed reasonable to remove anyone who had spent less than 4 minutes on the survey. However, this led to no further exclusions, given that the minimum time spent on the survey by any one participant was 6 minutes. The average time spent on the survey by the 70 remaining participants was 887 seconds (SD = 389). Forty-five of these participants were female and 25 male, and they had a mean age of 36 (SD = 11). Sixty-four indicated college as their highest education level, 3 high school, and 3 a lower education level.

Materials

We used the same six scenarios as in the previous experiments. Each scenario always appeared with two possible explanations, one being the target explanation from the second experiment (which was also the explanation from the strong condition of Experiment 1) and the other being either the strong or the intermediate alternative from the second experiment. In particular, the odd-numbered scenarios in the Appendix were shown with the strong alternative while the even-numbered scenarios where shown with the intermediate one. Each scenario appeared twice during the experiment, once with questions asking for the quality of the target and the alternative explanation, and once with a question asking whether the participant agreed with the target explanation, with "yes" and "no" as the two response options, which now was immediately followed by a question asking how confident the participant felt about his or her answer to the agreement question. The wording of this "metacognitive confidence" question was taken from Thompson, Prowse Turner, and Pennycook (2011): "In providing my answer to the above question I felt: . . ." Underneath the question appeared a Visual Analogue Scale, from 0 to 100, with "Guessing," "Fairly certain," and "Certain I'm right" appearing at the left end, in the middle, and at the right end of the scale, respectively.[12] For reasons of consistency, we also used Visual Analogue Scales with the same range for the explanation-quality questions. Here the left (0), middle (50), and right (100) positions were labeled "Very bad," "Neither good nor bad," and "Very good," respectively.

We used additional materials to distract the participants between answering the questions about explanation quality, on the one hand, and those about agreement and metacognitive confidence, on the other. Specifically, there was an essay question that asked participants to comment

[12] One could consider using a slightly different scale here, notably, one ranging from "Certain I'm wrong" to "Certain I'm right." However, given that as part of the initial screening we asked participants whether they had responded seriously (a question that all answered in the positive, as mentioned previously), we believe that as the lower end of the scale one best takes "Guessing"; anything "below" that point would indicate non-serious responding or even deceit.

on current wildlife crime and there were two questions asking participants about two pictures that they were shown. These questions did double duty as attention checks.

Design and procedure

All participants saw the same questions. The experiment consisted of three blocks: one block in which all scenarios were presented and in which participants were asked to rate the quality of the target and the alternative explanations; a second block in which also all scenarios were presented and in which participants were asked whether they agreed with the target explanation and how confident they felt in answering the question of agreement; and a third block, which contained various demographic questions as well as the materials mentioned earlier to distract the participants. The third block always appeared between the other two. The order in which those other blocks appeared was counterbalanced between participants, as was the order of the scenarios within those blocks. The purpose of the middle block was to reduce carry-over effects from the block that appeared first to the block that appeared last.

The scenarios were shown individually on screen, with the questions pertaining to them (questions concerning explanation quality, or questions of agreement and of metacognitive confidence) appearing on the same screen.

Results

The main aim of the present experiment was twofold: (i) to investigate the relation between, on the one hand, participants' judgments of the quality of both the target and the alternative explanations and, on the other hand, their agreement with the respective target explanations; and (ii) to investigate the relation between those judgments and participants' metacognitive confidence in their responses to the agreement questions.

So far, all results concerning the effect of judgments of explanation quality on agreement were carried out on the basis of aggregate data: percentages agreement and mean ratings of explanation quality (averaged across participants) for each item. Because the third experiment elicited participants' responses concerning both agreement and explanation quality, it gave us

the opportunity to regress the former on the latter and thereby gain insight into how the two hang together at the individual level. Figure 3 is a graphical summary of the data relevant to the regressions, plotting participants' ratings of explanation quality against their answers to the agreement questions for corresponding scenarios.

To investigate the effect of perceived explanation quality on participants' willingness to agree with the target explanation, we fitted three binomial GLMMs. All three models had the yes/no responses to the agreement questions as dependent variable and participant and item as random effect factors, using the maximal random-effects structure (Barr, Levy, Scheepers, & Tily, 2013). One model had quality ratings of the target explanation as fixed effect (this model is referred to as "GLMT"), one model had quality ratings of the target explanation and quality ratings of the alternative explanation as fixed effects (GLMTA), and one model had the difference between those ratings as fixed effect (GLMD). The R package `BayesFactor` supports the use of random effects as well, which allowed us to fit, for each of the aforementioned models, a Bayesian model with the same fixed- and random-effects structure.

Table 8 displays the model comparison results. There is controversy in the literature about which model comparison criteria to rely on in the case of logistic regression models, so we report

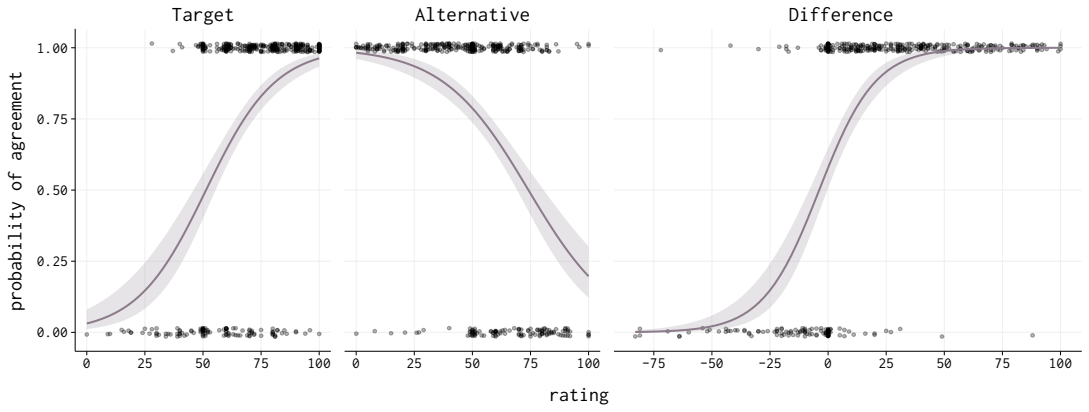

Figure 3: Comparing participants' agreement with the various target explanations to their quality ratings for those explanations (left) and to the quality ratings for the alternative explanations (middle), as well as to the difference of those ratings (right). Data are plotted with jitter to enhance visibility; smoothers have been added to highlight the trends (shaded bands indicate 95-percent-confidence intervals).

Table 8: Comparison of GLMMs from Experiment 3.

	k	LL	AIC	ΔAIC	BIC	ΔBIC	Count	R^2	D	AUC	BF
GLMTA	15	−132.79	295.57	1.51	356.18	29.80	.93	.46	.69	.89	1.7×10^{36}
GLMT	8	−182.71	381.42	87.36	413.73	87.35	.85	.26	.39	.77	2.5×10^{23}
GLMD	8	−139.03	294.06	0.00	326.38	0.00	.90	.43	.58	.85	2.5×10^{36}

Note: Count is the proportion of yes/no responses that the models classify correctly; R^2 gives the value of McFadden's pseudo-R^2; the number D is Tjur's coefficient of discrimination, which here equals the mean predicted value for the positive responses minus the mean predicted value for the negative responses; and AUC is the area under the Receiver Operating Curve, which quantifies how much better or worse is the trade-off between the true-positive rate and the false-positive rate of the model, compared to the same trade-off for mere guessing, averaged over all threshold values for qualifying as positive. NB: The BF value for a given GLMM pertains to the Bayesian model with the same fixed- and random-effects structure as the GLMM.

all standards ones.[13] It is clear that, across these criteria, the model that has only quality ratings of target explanation as fixed effect does worse than the other models, which take into account also quality ratings of alternative, either directly (GLMTA) or indirectly, via the difference between quality ratings of target explanation and of alternative explanation (GLMD). In terms of BIC, GLMD is the clear winner; in terms of AIC, too, it is the winner, although here GLMTA is a close second. In addition, GLMD comes out of the Bayesian analysis as the one most strongly favored by the data. The left panel of Figure 5 plots this model.

Table 9: Regression results for GLMTA and GLMD.

		b	SE	z	p
GLMTA	Intercept	0.34	1.47	0.23	.815
	T	0.19	0.03	6.18	< .0001
	A	−0.18	0.03	−7.08	< .0001
GLMT	Intercept	−4.34	0.79	−5.47	< .0001
	T	0.09	0.02	5.57	< .0001
GLMD	Intercept	0.31	0.20	1.56	.119
	D	0.14	0.02	7.72	< .0001

Note: T and A are the quality ratings of the target and alternative explanations, respectively; D is the vector of differences between those ratings.

[13] However, there is increasing consensus that, for model *selection*, AIC and/or BIC are to be relied on, as these give information about the models' predictive accuracy, that is, about how well they can be expected to fit *future* data; the other measures only indicate how well the models fit the *current* data. Models that provide excellent fit of the current data may do so due to overfitting, as a result of which they may do quite poorly in accounting for new data.

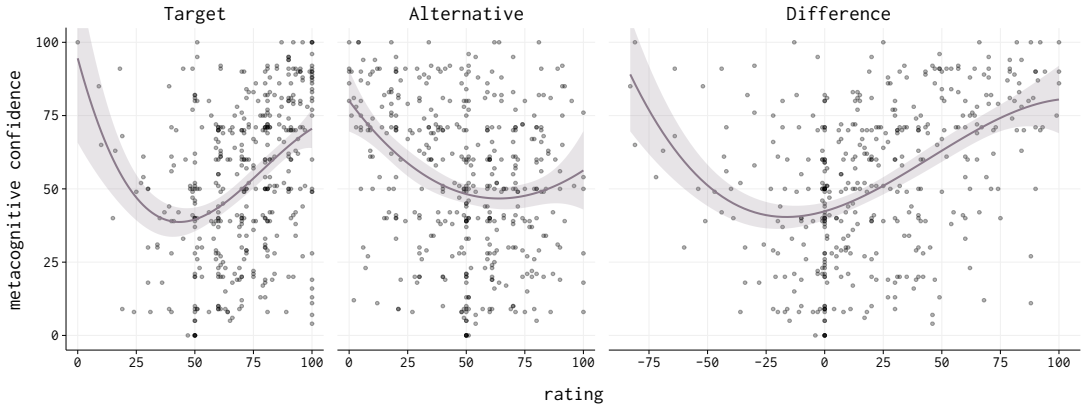

Figure 4: Comparing responses to metacognitive-confidence questions with judgments of explanation quality of target (left) and alternative (middle) explanation, and of the difference between the two (right), with smoothers added to highlight trends (shaded bands indicate 95-percent-confidence intervals).

Furthermore, it is encouraging to see that, for the best two models, the values for Count, McFadden's (1979) pseudo-R^2, Tjur's (2009) D, and the AUC all count as very good by conventional standards. (See the note to Table 8 for an explanation of the various measures.) And from a Bayesian perspective, the support for all three models is extremely strong.

Table 9 gives the regression results for the three models. From the coefficient for T in GLMTA, we derive an odds ratio of $\exp(0.19) \approx 1.21$, which indicates that for every percent of increase in the rating of the target explanation, the odds of agreement increase by (approximately) 21 percent, ceteris paribus; on the other hand, the odds ratio for A is $\exp(-0.18) \approx 0.83$, indicating that for every percent of increase in the rating of the alternative explanation, the odds of agreement with the target explanation decrease by (approximately) 17 percent, ceteris paribus. From the results for GLMD, we similarly derive that, for each percent of increase in the difference between ratings, there is a 15 percent increase in odds of agreement.

Our analysis of the effect of perceived explanation quality on metacognitive confidence almost exactly paralleled the above analysis of the relation between perceived explanation quality and agreement. Figure 4 gives a summary of the data relevant to the new analysis, comparing the ratings of explanation quality with participants' responses to the metacognitive-confidence questions for the same scenarios. Here, we fitted a number of LMMs. These models all had the responses to the metacognitive-confidence questions as dependent variable and also had partic-

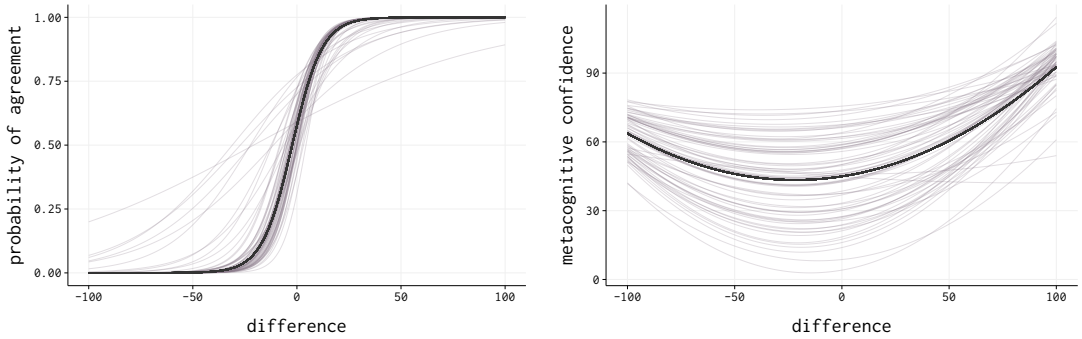

Figure 5: Predictions of GLMD for probability of agreement on the basis of difference in explanation quality ratings (left), and predictions of LMDQ for metacognitive confidence on the same basis (right). In both plots, the gray lines show the predictions of the individual random effects from all 70 participants.

ipant and item as random effects, again with the maximal random-effects structure. We fitted three models—referred to as "LMT," "LMTA," and "LMD"—that had the same fixed effects as GLMT, GLMTA, and GLD, respectively. Because Figure 4 suggested a quadratic trend in the effect on metacognitive confidence of the ratings of quality of the target explanation, and also of the difference between the ratings for the two explanations, we fitted three additional LMMs, which were like LMT, LMTA, and LMD except that they included a quadratic term for ratings of quality of the target explanation (in the case of LMT and LMTA) or a quadratic term for difference between ratings of quality of target and alternative explanation; these models are referred to as "LMTQ," "LMTAQ," and "LMDQ," respectively.

Table 10 presents the results from the model comparisons. LMDQ is far superior to the other models in terms of AIC and BIC. That the quadratic component in LMDQ significantly

Table 10: Comparison of LMMs from Experiment 3.

	k	LL	AIC	ΔAIC	BIC	ΔBIC	R^2	BF
LMTA	16	−474.15	980.30	21.64	1044.94	45.89	.75	4.1×10^{10}
LMT	9	−485.75	989.50	30.85	1025.86	26.80	.70	2.3×10^{8}
LMD	9	−480.47	978.93	20.28	1015.29	16.24	.71	7.5×10^{10}
LMTAQ	17	−467.82	969.64	10.99	1038.32	39.27	.76	1.9×10^{15}
LMTQ	10	−481.50	982.99	24.34	1023.40	24.34	.70	9×10^{12}
LMDQ	10	−469.33	958.65	0.00	999.06	0.00	.71	1.4×10^{18}

Note: For explanation, see the note to Table 1. The R^2-values were calculated using the `r2` function in the `sjstats` package (Lüdecke, 2017) for R, which follows the recommendations of Nakagawa and Schielzeth (2013). NB: The BF values pertain to the Bayesian models with the same fixed- and random-effects structure as the LMMs.

Table 11: Regression results for LMDQ.

		b	SE	β	*t*	*p*
LMDQ	Intercept	0.01	0.10		0.14	.89
	D1	6.93	1.00	0.34	6.93	< .0001
	D2	4.79	0.87	0.23	5.53	< .0001

Note: D1 is the linear component of difference between explanation-quality ratings, and D2 is the quadratic component of that difference. Standardized coefficients were obtained via the `std_beta` function in the `sjstats` package.

contributes to model fit was confirmed by a likelihood-ratio test comparing that model with LMD, its variant without the quadratic component: $\chi^2(1) = 21.85$, $p < .0001$. Further confirmation came from the regression results for LMDQ, as displayed in Table 11, which show that the quadratic component is highly significant.[14] The right panel of Figure 5 shows that the model closely captures the trend in the data that was suggested by our exploratory data analysis.

We here also briefly return to Lipton's idea that best explanations must be good enough to be acceptable, an idea for which we found only weak support in Experiment 1. Suppose there is, as per Lipton's idea, a threshold for explanation quality, which may be different for each participant, such that the participant agrees with a best explanation if, and only if, that explanation exceeds the threshold. Now that the same participants have both rated the quality of each explanation and indicated with which explanations they agreed, we should be able to find evidence for (or against) Lipton's proposal by looking whether we can associate with each participant at least one number such that the target explanations he/she deemed best and whose quality he/she rated above that number were accepted by him/her while the other target explanations were rejected. To verify this, we fitted a generalized linear model (GLM) for each participant individually, where the participant's responses to the agreement questions served as dependent variable and his/her quality ratings of those target explanations that he/she had deemed best served as the independent variable. We then calculated, for each resulting model, the proportion of responses to the agreement question that it classifies correctly, the so-called Count. A Count of 1

[14] For the purposes of model comparison, the LMMs were fit using the maximum-likelihood-estimation method, whereas for parameter estimation they were fit using restricted maximum-likelihood estimation; see, e.g., Pinheiro and Bates (2000, Ch. 2) for explanation.

indicates a perfect split in the responses on the basis of explanation-quality ratings, in the way just indicated: we can then perfectly predict whether a best explanation was accepted strictly on the basis of whether that explanation was given a rating above a certain threshold. For one participant, we could not fit a GLM because he or she had given the same (negative) response to all agreement questions. Of the GLMs fitted for the other 69 participants, 61 (or 88 percent) made no classification error, and only 2 made more than 2 classification errors. Averaged over the models, the Count was .96 (SD = 0.13).

Discussion

Our analysis concerned the relation between a person's judgment of the quality of an explanation and her willingness to agree with that explanation, as well as the relation between the former and her metacognitive confidence in her (dis)agreement. We found strong support for the existence of those relations.

Specifically, there is good reason to believe that, ceteris paribus, a person is more willing to agree with an explanation the better she judges that explanation to be, and also that, again ceteris paribus, she is more willing to agree with the explanation the poorer she judges the second-best available explanation to be. In fact, the best supported predictor of our participants' willingness to agree with a target explanation turned out to be the difference between their rating of the target explanation and their rating of the alternative. This finding closely parallels the finding of Douven and Schupbach (2015a) that the same difference in ratings is a reliable predictor of people's probability updates.

Toward the end of the previous section, we hypothesized that the drop in agreement with an explanation that is caused by staging a second explanation had itself a metacognitive explanation. Specifically, the thought was that, "best explanation" being an informal and vague notion, the introduction of an alternative could make people hesitant about inferring to the best explanation simply because they come to doubt their judgment as to which of the explanations is better. Such doubt would seem the likelier to arise the closer the two explanations are in terms

of perceived quality, which would explain why the drop in agreement tends to be absent, or in any case much less pronounced, when the alternative explanation is of noticeably lower quality than the target explanation. We found evidence for this hypothesis in the form of a significant effect of difference in ratings between target and alternative on participants' confidence in their responses to the agreement questions. When participants judged the two explanations to be about equally good, their metacognitive confidence tended to be low, while they became more confident the more they deemed the explanations to differ in quality.

Finally, we found support for the hypothesis that best explanations need to be good enough in order to be acceptable. For the vast majority of participants, we could perfectly predict whether they agreed with what they deemed the best explanation, based on whether or not, in their judgment, that explanation surpassed a certain threshold of goodness.

General discussion

Summary of main findings

Across three experiments, we found evidence that the quality of an explanation is a good predictor of people's willingness to agree with that explanation. Accordingly, the answer to question Q1 stated in the introduction is "yes." The first two experiments also showed that, while an explanation's probability was a good predictor of willingness to agree with the explanation, it was as such inferior to explanation quality.

The results from Experiment 2 (A and B), in conjunction with those from Experiment 1, showed that it matters for people's willingness to infer the truth of an explanation whether that explanation appears alone or side by side with a competitor, although it does *not* matter for how they perceive the quality of the explanation, thus answering the first part of Q2 in the negative and the second part in the positive. The same results showed that the quality of the competitor matters to people's willingness to infer the truth of the target explanation: if the competitor is about as good as the target explanation, people tend to be significantly less inclined to infer the

truth of the target explanation, whereas the effect of introducing a second explanation tends to be smaller when the competitor is clearly inferior to the target explanation. That is a "yes" to Q3.

The data from Experiment 3 offered the most detailed answers to Q2 and Q3, in that participants' responses to the agreement questions could be reliably predicted from the difference between their judgment of the quality of the target explanation and that of the quality of the alternative. From this, we gained a deeper understanding of exactly how the presence of a second candidate explanation, as well as its quality in comparison with that of the target explanation, is related to people's willingness to infer to the best explanation. The same experiment also gave some insight into *why* the presence and quality of a second explanation matter: if the explanations are about equally good, that undermines people's confidence in the inference. This tendency is understandable, given that there is no known formal criterion that one could apply as a kind of litmus test to determine, in case of doubt, which of the two explanations is really better.

Douven and Schupbach (2015a) had shown that people's probability updates are guided by explanatory considerations. However, their studies were limited by the choice of their materials, which concerned a somewhat artificial setting in which hypotheses could be said to explain the evidence only in a shallow sense. Also, they only considered probability updates and not the textbook version of IBE or its amended cousins. There was no previous reason to hold that the connection between explanatory power and what people are willing to infer could be systematically described. Our experiments give compelling evidence that the connection *is* systematic and is governed by a rule that has been proposed in the philosophical literature on normative grounds.

The results are also consistent with those reported in Tenney, Cleary, and Spellman (2009), which showed that introducing a second, equally plausible suspect significantly lowered the number of "guilty" verdicts reached by their participants. More importantly, we can now answer some important questions left open by that previous research. Specifically, Tenney et al. left for future work varying the initial plausibility of the alternative suspect. Based on our results, we can

safely predict that this would affect the drop in "guilty" verdicts, even to the extent that introducing a less plausible alternative suspect might not affect the proportion of "guilty" verdicts at all. More interestingly, we now have an understanding of *why* the drop in "guilty" verdicts occurred in Tenney et al.'s study: people's inferences are sensitive to explanatory considerations, and by introducing an alternative suspect we present participants with a rival explanation. Experiment 3 gives even more insight on this point: depending on exactly how good the rival explanation is, people's metacognitive confidence in the explanatory status of the guilt of the primary suspect may decrease, possibly enough to block the inference to the guilt of that suspect.

Directions for future research

In the introduction, we briefly mentioned versions of IBE that qualify the textbook version in ways different from, and possibly complementary to, Bird's and Lipton's proposals. A particularly promising proposal replaces the inference to the *truth* of the best explanation with the inference to the best explanation's being *closer* to the truth than its known rivals (see Douven, 2002, 2017a). Here, too, one might require that the best explanation be good enough, and also that it be considerably better than its closest competitor. Our materials did not lend themselves to investigating this version of IBE. The notion of truth-closeness does not make sense for all hypotheses (Niiniluoto, 1998): we know what is meant when we are told that Newtonian mechanics is, while false, still close to the truth, and closer to the truth than any of its predecessors; it is much more difficult to make sense of the claim that, although Mrs. Smith was actually murdered by her husband, the hypothesis that she was murdered by her colleague Mr. Hanson is still close to the truth. Nevertheless, it seems worthwhile to extend the current work to versions of IBE that refer to truth-closeness instead of truth simpliciter.

A second avenue for further research concerns the connection Experiment 3 made with the literature on metacognition, as developed in Ackerman (2014), Ackerman and Thompson (2015, 2017a, 2017b), Thompson, Prowse Turner, and Pennycook (2011), and other publications. The full experimental paradigm, as developed in Thompson, Prowse Turner, and Pennycook

(2011), consists of two stages, one in which participants are asked for a quick first response to whatever the task at hand is, where this response is immediately followed by a confidence rating, and a second stage in which participants are asked for a second response to the same task but are now given more time to answer, and are then again asked how confident they were in responding. This procedure is meant to shed light on which factors make participants engage in effortful ("type 2") thinking rather than stick to their quick and effortless ("type 1") first response (Evans, 2007; Kahneman, 2011). The metacognitive part of Experiment 3 consisted of only one stage. While the results warranted the conclusion that difference in ratings of quality of best and second-best explanation has a metacognitive effect, it would be interesting to employ in future research the full metacognitive paradigm, which could provide more detailed information about when differences in quality ratings give rise to effortful thinking.

Furthermore, in Experiments 2 (A and B) and 3 we introduced only one alternative explanation. Although Tenney, Cleary, and Spellman's (2009) study gives reason to expect that introducing further alternative explanations would lead to results very similar to the ones reported here, it is still worth verifying this explicitly. Having participants consider various competing explanations would also make the investigation more relevant to how IBE is used in actual scientific practice, where researchers often have multiple hypotheses on the table and hope to find out which of them, if any, is true. In addition to this, we plan a study in which participants are first asked to generate possible explanations themselves—rather than those being presented to them—and then to rate the quality of the explanations they have come up with.[15]

We looked at a normative proposal from a descriptive perspective. The finding that people not only tend to infer to the best explanation, but in doing so also attend to the difference in quality between best and second-best explanation, makes it interesting to investigate what the *normative* status of that practice is. Bird and Lipton have made their proposals look intuitively appealing, but that is not the same as showing that it is indeed epistemically helpful to give weight in one's reasoning to the difference in quality between the best and the second-best explanation.

[15] This idea was inspired by comments from an anonymous referee, for which we thank him/her.

For all they have said, we might still be better advised to ignore that difference. Douven and Schupbach's (2015b) finding that some measures of explanatory power do better in predicting probability updates than others has led to new normative results concerning probabilistic versions of IBE that base their attribution of an explanation bonus on a measure of explanatory power.

For example, Douven (2017b) uses computer simulations to study various instances of the update schema

$$\Pr[\![E]\!](H_i) \;=\; \frac{\Pr(H_i)\Pr(E \mid H_i) + c\Pr(H_i)\Pr(E \mid H_i)\mathscr{M}(H_i, E)}{\sum_{j=1}^{n}\big(\Pr(H_j)\Pr(E \mid H_j) + c\Pr(H_j)\Pr(E \mid H_j)\mathscr{M}(H_j, E)\big)}, \qquad \text{(US)}$$

with Pr one's subjective probability function over a given hypothesis partition $\mathscr{H} = \{H_1, \ldots, H_n\}$ prior to learning evidence E; $\Pr[\![E]\!](H_i)$ one's new subjective probability for hypothesis H_i immediately after learning E (and nothing stronger); $\mathscr{M}(H_j, E)$ a measure of the explanation quality of H_i in light of E (how well does H_i explain E); and $c \in [0, 1]$ a constant determining what percentage of H_i's probability is added in proportion to this hypothesis' power to explain E. Most prominent among the measures of explanation quality Douven considers are Popper's (1959) measure, according to which the explanatory goodness of an H_i in light of E equals

$$\frac{\Pr(E \mid H_i) - \Pr(E)}{\Pr(E \mid H_i) + \Pr(E)},$$

and Good's (1960) measure,[16]

$$\ln\left(\frac{\Pr(E \mid H_i)}{\Pr(E)}\right).$$

In the computer simulations, updating via the instances of (US) with either of the above measures came out as being superior to updating via Bayes' rule—the instance of (US) with $c = 0$—in

[16] Actually, Douven considers a particular functional rescaling of Good's measure, called "L_2" in Douven and Schupbach (2015b); the details need not detain us here.

a variety of important respects, notably, speed of convergence to the objectively correct probability and overall accuracy.

In light of the findings reported in this paper, however, it would make sense to consider variants of (US) that replace $\mathscr{M}(H_i, E)$ by a measure we might designate as $\mathscr{D}(\mathscr{H}, E)$, which specifically compares the explanation quality of the best and second-best explanation among the hypotheses in $\mathscr{H}$ and bases the possible "explanation bonus" also, or perhaps even strictly, on the *difference* in quality between best and second-best explanation. We intend to investigate versions of $\mathscr{D}(\mathscr{H}, E)$ from a normative perspective in future research.

Finally, we come back to a point touched upon in the theoretical background section, viz., that the finding that explanatory considerations play a role in people's inferential behavior in ways that deviate from Bayes' rule is perfectly compatible with the key insights from New Paradigm psychology of reasoning (Over, 2009), according to which reasoning is essentially probabilistic. The update schema (US) may serve to buttress this point, inasmuch as, supposing $\mathscr{M}$ to be Popper's or Good's measure or some other probabilistic measure of explanation quality, the resulting update rule will, for any choice of the constant c, result in a *strictly* probabilistic update rule. It is not difficult to think of explications of $\mathscr{D}(\mathscr{H}, E)$ above that make this measure, like Popper's and Good's, a strictly probabilistic measure.

Acknowledgments

We are greatly indebted to Rakefet Ackerman, Mike Oaksford, Christopher von Bülow, Michael Waldmann, and an anonymous referee for valuable comments on previous versions of this paper. We owe a special debt to Henrik Singmann for helpful statistical advice. We are also grateful to audiences in Aurich, London, Munich, Norwich, and Paris for helpful questions and remarks.

References

Ackerman, R. (2014). The diminishing criterion model for metacognitive regulation of time investment. *Journal of Experimental Psychology: General*, *143*, 1349–1368.

Ackerman, R., & Thompson, V. A. (2015). Meta-reasoning: What we can learn from meta-memory. In A. Feeney & V. A. Thompson (eds.), *Reasoning as memory* (pp. 164–178). Hove UK: Psychology Press.

Ackerman, R., & Thompson, V. A. (2017a). Meta-reasoning: Shedding meta-cognitive light on reasoning research. In L. Ball & V. A. Thompson (eds.), *International handbook of thinking and reasoning*. Hove UK: Psychology Press, in press.

Ackerman, R., & Thompson, V. A. (2017b). Meta-reasoning: Monitoring and control of thinking and reasoning. *Trends in Cognitive Sciences*, *21*, 607–617.

Ali, N., Chater, N., & Oaksford, M. (2011). The mental representation of causal conditional reasoning: Mental models or causal models. *Cognition*, *119*, 403–418.

Aust, F., Diedenhofen, B., Ullrich, S., & Musch, J. (2013). Seriousness checks are useful to improve data validity in online research. *Behavior Research Methods*, *45*, 372–400.

Baayen, H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412.

Baillargeon, R., & DeJong, G. F. (2017). Explanation-based learning in infancy. *Psychonomic Bulletin & Review*, in press.

Baratgin, J., & Politzer, G. (2007). The psychology of dynamic probability judgment: Order effect, normative theories, and experimental methodology. *Mind & Society*, *6*, 53–66.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255–278.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). `lme4`*: Linear mixed-effects models using eigen and S4*. R package version 1.1–0 (available at `http://lme4.r-forge.r-project.org/`).

Bes, B., Sloman, S., Lucas, C. G., & Raufaste, E. (2012). Non-Bayesian inference: Causal structure trumps correlation. *Cognitive Science*, *36*, 1178–1203.

Bird, A. (2010). Eliminative abduction: Examples from medicine. *Studies in the History and Philosophy of Science*, *41*, 345–352.

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

Boyd, R. (1981). Scientific realism and naturalistic epistemology. In P. Asquith & R. Giere (eds.), *Proceedings of the PSA 1980* (vol. II, pp. 613–662). East Lansing MI: Philosophy of Science Association.

Bunt, H. C., & Black, W. J. (2000). *Abduction, belief, and context in dialogue: Studies in computational pragmatics*. Cambridge MA: MIT Press.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multi-model inference: A practical information-theoretic approach*. Berlin: Springer.

Byrne, M. D. (1995). The convergence of explanatory coherence and the story model: A case study in juror decision. In *Proceedings of the seventeenth annual conference of the Cognitive Science Society* (pp. 539–543). Mahwah NJ: Erlbaum.

Chi, M. T. H., de Leeuw, N., Chiu, M.-H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, *18*, 439–477.

de Finetti, B. (1937). Foresight: Its logical laws, its subjective sources. In H. Kyburg & H. Smokler (eds.), *Studies in subjective probability* (pp. 53–118). New York: Krieger (1980).

Douven, I. (1999). Inference to the best explanation made coherent. *Philosophy of Science*, *66*, S424–S435.

Douven, I. (2002). Testing inference to the best explanation. *Synthese*, *130*, 355–377.

Douven, I. (2013). Inference to the best explanation, Dutch books, and inaccuracy minimisation. *Philosophical Quarterly*, *69*, 428–444.

Douven, I. (2016a). *The epistemology of indicative conditionals*. Cambridge: Cambridge University Press.

Douven, I. (2016b). Explanation, updating, and accuracy. *Journal of Cognitive Psychology*, *28*, 1004–1012.

Douven, I. (2017a). Abduction. In E. N. Zalta (ed.), *Stanford encyclopedia of philosophy*, https://plato.stanford.edu/archives/summer2017/entries/abduction/.

Douven, I. (2017b). Inference to the best explanation: What is it? And why should we care? In T. Poston & K. McCain (eds.), *Best explanations: New essays on inference to the best explanation* (pp. 4–22). Oxford: Oxford University Press.

Douven, I., Elqayam, S., Singmann, H., & van Wijbergen-Huitink, J. (2018). Conditionals and inferential connections: A hypothetical inferential theory. *Cognitive Psychology*, in press.

Douven, I., & Meijs, W. (2007). Measuring coherence. *Synthese*, *156*, 405–425.

Douven, I., & Schupbach, J. N. (2015a). The role of explanatory considerations in updating. *Cognition*, *142*, 299–311.

Douven, I., & Schupbach, J. N. (2015b). Probabilistic alternatives to Bayesianism: The case of explanationism. *Frontiers in Psychology*, *6*, doi: 10.3389/fpsyg.2015.00459.

Douven, I., & Wenmackers, S. (2017). Inference to the best explanation versus Bayes' rule in a social setting. *British Journal for the Philosophy of Science*, *68*, 535–570.

Evans, J. St. B. T. (2007). *Hypothetical thinking: Dual processes in reasoning and judgement*. Hove UK: Psychology Press.

Fernbach, P. M., Darlow, A., & Sloman, S. A. (2010). Neglect of alternative causes in predictive but not diagnostic reasoning. *Psychological Science*, *21*, 329–336.

Fernbach, P. M., & Erb, C. D. (2013). A quantitative causal model of conditional reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 1327–1343.

Fischoff, B., & Lichtenstein, S. (1978). Don't attribute this to Reverend Bayes. *Psychological Bulletin*, *85*, 239–243.

Fitelson, B. (2003). A probabilistic theory of coherence. *Analysis*, *63*, 194–199.

Good, I. J. (1960). Weight of evidence, corroboration, explanatory power, information and the utility of experiment. *Journal of the Royal Statistical Society*, *B22*, 319–331.

Hall, S., Ali, N., Chater, N., & Oaksford, M. (2016). Discounting and augmentation in causal conditional reasoning: Causal models or shallow encoding? *PLoS ONE*, *11*, e0167741, doi:10.1371/journal.pone.0167741.

Hemmerich, J. A., Van Voorhis, K., & Wiley, J. (2016). Anomalous evidence, confidence change, and theory change. *Cognitive Science*, *40*, 1534–1560.

Hobbs, J. R. (1992). Metaphor and abduction. In A. Ortony, J. Slack, & O. Stock (eds.), *Communication from an artificial intelligence perspective* (pp. 35–58). New York: Springer.

Hobbs, J. R. (2004). Abduction in natural language understanding. In L. Horn & G. Ward (eds.), *Handbook of pragmatics* (pp. 724–741). Oxford: Blackwell.

Holyoak, K. J., & Lee, H. S. (2017). Inferring causal relations by analogy. In M. R. Waldmann (ed.), *The Oxford handbook of causal reasoning* (pp. 459–473). Oxford: Oxford University Press.

Holyoak, K. J., Lee, H. S., & Lu, H. (2010). Analogical and category-based inference: A theoretical integration with Bayesian causal models. *Journal of Experimental Psychology: General*, *139*, 702–727.

Horry, R., & Brewer, N. (2016). How target–lure similarity shapes confidence judgments in multiple-alternative decision tasks. *Journal of Experimental Psychology: General*, *145*, 1615–1634.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*, 434–446.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford: Oxford University Press.

Joyce, J. (1998). A nonpragmatic vindication of probabilism. *Philosophy of Science*, *65*, 575–603.

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Macmillan.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Keil, F. C. (2006). Explanation and understanding. *Annual Review of Psychology*, *57*, 227–254.

Kitcher, P. (1981). Explanatory unification. *Philosophy of Science*, *48*, 507–531.

Lagnado, D. A., & Gerstenberg, T. (2017). Causation in legal and moral reasoning. In M. R. Waldmann (ed.), *The Oxford handbook of causal reasoning* (pp. 565–601). Oxford: Oxford University Press.

Lange, M. (2017). *Because without cause: Non-causal explanation in science and mathematics*. Oxford: Oxford University Press.

Legare, C. H., & Lombrozo, T. (2014). Selective effects of explanation on learning during early childhood. *Journal of Experimental Child Psychology*, *126*, 198–212.

Legare, C. H., Sobel, D. M., & Callanan, M. (2017). Causal learning is collaborative: Examining explanation and exploration in social contexts. *Psychonomic Bulletin & Review*, in press.

Lenth, R. (2015). `lsmeans`*: Least-squares means*. R package version 2.20–23, `http://cran.r-project.org/package=lsmeans`.

Lewis, D. (2000). Causation as influence. *Journal of Philosophy*, *97*, 182–197.

Lipton, P. (1993). Is the best good enough? *Proceedings of the Aristotelian Society*, *93*, 89–104.

Lipton, P. (2004). *Inference to the best explanation* (2nd ed.). London: Routledge.

Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences*, *10*, 464–470.

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology*, *55*, 232–257.

Lombrozo, T. (2016). Explanatory preferences shape learning and inference. *Trends in Cognitive Sciences*, *20*, 748–759.

Lombrozo, T., & Gwynne, N. Z. (2014). Explanation and inference: Mechanistic and functional explanations guide property generalization. *Frontiers in Human Neuroscience*, *8*, doi: 10.3389/fnhum.2014.00700.

Lombrozo, T., & Vasilyeva, N. (2017). Causal explanation. In M. R. Waldmann (ed.), *The Oxford handbook of causal reasoning* (pp. 415–432). Oxford: Oxford University Press.

Lüdecke, D. (2017). sjstats*: Statistical functions for regression models*. R package version 0.8.0 (available at https://CRAN.R-project.org/package=sjstats).

Mancosu, P. (2015). Explanation in mathematics. In E. N. Zalta (ed.), *Stanford encyclopedia of philosophy*, https://plato.stanford.edu/archives/sum2015/entries/mathematics-explanation/.

McFadden, D. (1979). Quantitative methods for analyzing travel behaviour of individuals: Some recent developments. In D. Hensher & P. Stopher (eds.), *Behaviour travel modelling* (pp. 279–318). London: Croom Helm.

McMullin, E. (1992). *The inference that makes science*. Milwaukee WI: Marquette University Press.

McMullin, E. (1996). Epistemic virtue and theory appraisal. In I. Douven & L. Horsten (eds.), *Realism in the sciences* (pp. 1–34). Leuven: Leuven University Press.

Meder, B., & Mayrhofer, R. (2017). Diagnostic reasoning. In M. R. Waldmann (ed.), *The Oxford handbook of causal reasoning* (pp. 433–457). Oxford: Oxford University Press.

Meder, B., Mayrhofer, R., & Waldmann, M. R. (2014). Structure induction in diagnostic causal reasoning. *Psychological Review*, *121*, 277–301.

Morey, R. D., & Rouder, J. N. (2015). BayesFactor*: Computation of Bayes factors for common designs*. R package version 0.9.11-1 (available at http://CRAN.R-project.org/package=BayesFactor).

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*, 133–142.

Niiniluoto, I. (1998). Verisimilitude: The third period. *British Journal for the Philosophy of Science*, *49*, 1–29.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality*. Oxford: Oxford University Press.

Oaksford, M., & Chater, N. (2013). Dynamic inference and everyday conditional reasoning in the new paradigm. *Thinking & Reasoning*, *19*, 346–379.

Oaksford, M., & Chater, N. (2017). Causal models and conditional reasoning. In M. R. Waldmann (ed.), *The Oxford handbook of causal reasoning* (pp. 327–346). Oxford: Oxford University Press.

Olsson, E. J. (2002). What is the problem of coherence and truth? *Journal of Philosophy*, *94*, 246–272.

Over, D. E. (2009). New paradigm psychology of reasoning. *Thinking & Reasoning*, *15*, 431–438.

Pacer, M, Williams, J. J., Chen, X., Lombrozo, T., & Griffiths, T. L. (2013). Evaluating computational models of explanation using human judgments. In A. Nicholson & P. Smyth (eds.), *Proceedings of the twenty-ninth conference on uncertainty in artificial intelligence* (pp. 498–507). Corvallis OR: AUAI press.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Mateo CA: Morgan Kaufmann.

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

Pennington, N., & Hastie, R. (1986). Evidence evaluation in complex decision making. *Journal of Personality and Social Psychology*, *51*, 242–258.

Pennington, N., & Hastie, R. (1992). Explaining the evidence: Tests of the story-model for juror decision making. *Journal of Personality and Social Psychology*, *62*, 189–206.

Pennycook, G., Trippas, D., Handley, S. J., & Thompson, V. A. (2014). Base rates: Both intuitive and neglected. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*, 544–554.

Pinheiro, J., & Bates, D. (2000). *Mixed-effects models in S and S-Plus*. New York: Springer.

Popper, K. R. (1959). *The logic of scientific discovery*. London: Hutchinson.

Psillos, S. (2004). Inference to the best explanation and Bayesianism. In F. Stadler (ed.), *Induction and deduction in the sciences* (pp. 83–91). Dordrecht: Kluwer.

Putnam, H. (1975). *Mathematics, matter and method* (Philosophical papers, vol. I). Cambridge: Cambridge University Press.

Ramsey, F. P. (1926). Truth and probability. In his *Foundations of mathematics* (pp. 156–198). London: Routledge (1931).

R Core Team (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, http://www.R-project.org/.

Rittle-Johnson, B., & Loehr, A. M. (2017). Eliciting explanations: Constraints on when self-explanation aids learning. *Psychonomic Bulletin & Review*, in press.

Robinson, L. B., & Hastie, R. (1985). Revision of beliefs when a hypothesis is eliminated from consideration. *Journal of Experimental Psychology: Human Perception and Performance*, *11*, 443–456.

Rosenkrantz, R. D. (1992). The justification of induction. *Philosophy of Science*, *59*, 527–539.

Schum, D. A., & Martin, A. W. (1982). Formal and empirical research on cascaded inference in jurisprudence. *Law and Society Review*, *17*, 105–151.

Shoots-Reinhard, B. L., Rucker, D. D., Petty, R. E., & Shakarchi, R. (2014). Not all contrast effects are created equal: Extent of processing affects contrast strength. *Journal of Applied Social Psychology*, *44*, 523–535.

Sidney, P. G., Hattikudur, S., & Alibali, M. W. (2015). How do contrasting cases and self-explanation promote learning? Evidence from fraction division. *Learning and Instruction*, *40*, 29–38.

Singmann, H., Bolker, B., Westfall, J., & Aust, F. (2017). `afex`*: Analysis of factorial experiments*. R package version 0.17-8 (available at https://cran.r-project.org/package=afex).

Sloman, S. A. (1997). Explanatory coherence and the induction of properties. *Thinking & Reasoning*, *3*, 81–110.

Sloman, S. A. (2005). *Causal models: How we think about the world and its alternatives*. Oxford: Oxford University Press.

Stroup, W. W. (2012). *Generalized linear mixed models: Modern concepts, methods and applications*. Boca Raton FL: CRC Press.

Teller, P. (1973). Conditionalization and observation. *Synthese*, *26*, 218–258.

Tenney, E. R., Cleary, H. M. D., & Spellman, B. A. (2009). Unpacking the doubt in 'beyond a reasonable doubt': Plausible alternative stories increase not guilty verdicts. *Basic and Applied Social Psychology*, *31*, 1–8.

Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences*, *12*, 435–502.

Thompson, V. A., Prowse Turner, J. A., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology*, *63*, 107–140.

Tjur, T. (2009). Coefficients of determination in logistic regression models—a new proposal: The coefficient of discrimination. *American Statistician*, *63*, 366–372.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124–1131.

Vasilyeva, N., Wilkenfeld, D., & Lombrozo, T. (2017). Contextual utility affects the perceived quality of explanations. *Psychonomic Bulletin & Review*, in press.

Vogel, J. (1998). Inference to the best explanation. In E. Craig (ed.), *Routledge encyclopedia of philosophy*, London: Routledge (available at https://www.rep.routledge.com/articles/thematic/inference-to-the-best-explanation/v-1).

Waldmann, M. R. (2000). Competition among causes but not effects in predictive and diagnostic learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 53–76.

Waldmann, M. R., & Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *Journal of Experimental Psychology: General*, *121*, 222–236.

Walker, C. M., & Lombrozo, T. (2017). Explaining the moral of the story. *Cognition*, *167*, 266–281.

Wiegmann, A., & Waldmann, M. R. (2014). Transfer effects between moral dilemmas: A causal model theory. *Cognition*, *131*, 28–43.

Wilkenfeld, D. A., & Lombrozo, T. (2015). Inference to the best explanation (IBE) versus explaining for the best inference (EBI). *Science & Education*, *24*, 1059–1077.

Williams, J. J., & Lombrozo, T. (2010). The role of explanation in discovery and generalization: Evidence from category learning. *Cognitive Science*, *34*, 776–806.

Williams, J. J., & Lombrozo, T. (2013). Explanation and prior knowledge interact to guide learning. *Cognitive Psychology*, *66*, 55–84.

Williamson, T. (2007). *The philosophy of philosophy*. Oxford: Blackwell.

Wouters, A. (1995). Viability explanation. *Biology & Philosophy*, *10*, 435–457.

Wouters, A. (2007). Design explanation: Determining the constraints on what can be alive. *Erkenntnis*, *67*, 65–80.

Zemla, J. C., Sloman, S., Bechlivanidis, C., &. Lagnado, D. A. (2017). Evaluating everyday explanations. *Psychonomic Bulletin & Review*, in press.

Appendix

Each scenario has four possible explanations: The weak explanation was only used in Experiment 1. The strong explanation was used for the strong condition in that experiment, and it served as the target explanation for Experiments 2A, 2B, and 3. The strong alternative explanation was used in all experiments, serving as the intermediate condition in Experiment 1. The intermediate alternative explanation was used in all experiments except the first.

1. Mrs. Smith, a high-ranking administrator from a top-tier university, was found strangled in her office.

(Weak) One possibility is that she stumbled on the carpet as she was putting her scarf on and accidentally strangled herself to death. The scarf that she normally wears was found in her handbag and not around her neck.

(Strong) She was in the process of divorcing her husband, Mr. Smith, because she had fallen in love with another man and wanted to pursue this new relationship. Both she and her husband were seeking the custody of their two children. A surveillance video showed Mr. Smith leaving the building in which his wife's office is located approximately 30 minutes before Mrs. Smith's body was discovered. Mr. Smith had in the past also been accused of domestic violence, in particular connected with his very strong jealousy.

(Strong alternative) Mrs. Smith knew about *extremely incriminating evidence against Mr. Hanson*, one of her coworkers. The evidence is in fact so damaging that it *could lead to the termination* of Mr. Hanson's contract. Mr. Hanson was desperate to *keep Mrs. Smith from revealing his secret*. On the day of the murder, Mr. Hanson was *in his office, which is close* to Mrs. Smith's office.

(Intermediate alternative) Mr. Hanson, one of Mrs. Smith's coworkers, ***had a crush on her, but she had turned him down***. In fact, she had done so ***in a rather rude way***. Mr. Hanson had been depressed since, and had ***started drinking heavily***. In the past few weeks, Mrs. Smith had complained to Human Resources twice about Mr. Hanson ***insulting her and threatening her*** during work hours.

2. Lady Windermere was found murdered in her castle. A police investigation is under way to find the culprit.

(Weak) Lady Windermere's niece, who was visiting her during the week of the murder, likes to hunt. She was away the evening of the murder but left her hunting gun on the coffee table with the security on, but maybe the gun fired itself nonetheless and killed Lady Windermere.

(Strong) Her sister had the following motive: not only did the two sisters hate each other since they were children, the sister would also inherit Lady Windermere's fortune of £5,000,000 if the latter died, money that the sister badly needed for a life-saving operation. According to the coroner, Lady Windermere died at 8 PM and the sister was seen near Lady Windermere's castle 20 minutes before by a trustworthy witness.

*(Strong alternative // **Intermediate Alternative**)* Jeeves, Lady Windermere's butler, had the following motive: he owed Lady Windermere *£250,000* // ***£25,000***, which he had borrowed from her, and which *for him is an enormous amount of money* // ***which he would be able to reimburse within the next two years***. *He had also occasionally complained to his friends about Lady Windermere being too strict* // ***Also, Jeeves had always been very fond of Lady Windermere.*** *His wife was the only one who could testify that he spent the evening at home.* // ***His wife and a waiter testified that he spent the evening in a local restaurant, having dinner.***

3. Two weeks ago, a privately-owned Rembrandt was stolen from the house of its owner, John Brimmer II, the CEO of a multinational company.

(Weak) Mr. Brimmer's house is very large and it has a lot of rooms. The Rembrandt had recently been cleaned on-site by a specialized firm. The cleaning staff could have misplaced the painting and put it in a room that is rarely used, so that it hasn't been found yet.

(Strong) The Rembrandt had recently been cleaned on-site by a specialized firm. The police are currently focusing on an employee who had assisted with the cleaning and who suddenly quit his job last week. The police discovered that this employee had gambling debts totaling over $1,000,000 and that he had a criminal record for burglary and theft.

*(Strong alternative // **Intermediate Alternative**)* The police are aware of the fact that Mr. Brimmer had been experiencing significant financial difficulties lately, *even to the extent that he may have to file for bankruptcy soon* // ***although these were thought to be of no real concern for someone as wealthy as Mr. Brimmer***. *Also, the fact that there are no visible signs of burglary and the fact that the house's surveillance video system failed on the night the Rembrandt was stolen makes the police look seriously into the possibility that they are dealing with a case of insurance fraud.* // ***The fact that the house's surveillance video system failed on the night the Rembrandt was stolen suggests that this might be a case of insurance fraud, although the surveillance system is known to malfunction from time to time, for no apparent reason.***

4. You have radio contact with an isolated village, located in a valley below a dammed reservoir. The person at the other end says that the village is flooded, but when you ask her to elaborate, the radio contact breaks down.

(Weak) The village could have been flooded by a leaky faucet in one of the villagers' kitchen.

(Strong) You know that the dam near the village was recently checked, and that then several serious cracks were discovered in the dam. The report of the responsible authorities noted the state of the dam as alarming. You know also that the 4 engineers in charge of the dam are currently away for a training seminar in the capital. In the past 20 years, technical issues with the dam caused the village to be flooded 7 times.

*(Strong alternative // **Intermediate Alternative**)* There has also been rainfall in the valley for the past *7 days* // ***two days***. This is *a little unusual* // ***an usual occurrence*** for this time of year. The rain *has* // ***not*** been heavy and the earth in this area is generally *a bit hard* // ***quite soft*** so *it doesn't absorb the rain so well* // ***it absorbs the rain well.***

5. A team of scientists is collecting data about animal behavior in a remote part of the jungle. For this purpose, they installed an expensive camera in a wooden box and attached it as securely as possible to the trunk of a tree. The camera suddenly stopped sending pictures.

(Weak) Someone might have hiked across this remote part of the jungle, which is a hundred miles from the nearest village and which is inhabited by dangerous animals such as snakes and tigers, and that person could have stolen the box.

(Strong) A known problem with this type of camera is that the film is extremely sensitive to heat and humidity. On the day it broke, the temperature was 55˚C and the humidity was well above average, too. Also, the vent-holes of the box have often been found occluded by dirt and mud.

*(Strong alternative // **Intermediate Alternative**) Sometimes* // ***Rarely***, in previous experiments, similar cameras have been discovered and destroyed by curious animals. *The camouflage paint of the box tends to get scratched which makes it more conspicuous and the tree could have been shaken, making the box fall.* // ***However, the camouflage paint of the box was very realistic and***

it was strongly bolted to the tree so it is unlikely that it would have fallen even if the tree had been shaken. *A couple of monkeys* // ***Only a couple of small birds*** had been living on the neighboring *trees* // ***tree*** for the past days.

6. Ms. Hurley is at an appointment with her doctor. She has been having difficulty breathing, experiencing heavy night sweats and feeling constantly exhausted.
(Weak) One possibility is that Ms Hurley has a simple cold and has been sweating and sleeping badly at night because she might sleep with too many covers.
(Strong) She could suffer from pulmonary tuberculosis. During the appointment, she has been coughing a lot, and frequent coughing is normally a symptom of that disease. Moreover, she recently returned from Mali, a country where tuberculosis—a disease caused by a highly contagious bacteria—is endemic. Finally, she hasn't been vaccinated against tuberculosis.
(Strong alternative // ***Intermediate Alternative)*** Another possibility is that she suffers from lymphoma (a type of blood cell tumor). *She is currently being treated for arthrosis,* // ***In the past, she has been treated for arthrosis***, which *is a serious risk factor for developing lymphomas* // ***can be a risk factor for lymphomas***. Her lymph nodes are *mildly swollen* // ***very slightly swollen***: important swelling is normally a symptom of the disease. Finally, iron deficiency is often another symptom of lymphoma, and her iron levels are *below* // ***a little below*** the normal range.