Cognition 241 (2023) 105630

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Original articles

Revisiting the narrow latent scope bias in explanatory reasoning

Simon Stephan

Department of Psychology, University of Göttingen, Gosslerstrasse 14, 37073 Göttingen, Germany

ARTICLE INFO

Dataset link: https://simonstephan31.github.io/revisit_nlsbias/index.html

Keywords:
Narrow latent scope bias
Explanation
Causal reasoning
Diagnostic reasoning
Causal structure
Categorization
Pragmatics

ABSTRACT

Humans are capable explainers and lay people tend to share the same explanatory virtues held in high regard by philosophers and scientists. However, a recent line of studies found a striking deviation from normativity in lay people's explanations, termed the "narrow latent scope bias". When competing explanations with identical a priori probabilities fit observed evidence equally well – but differ in the number of unobserved pieces of evidence they predict (latent scope) – reasoners seem to prefer explanations that predict fewer unobserved pieces of evidence (narrow latent scope). This tendency has been described as a *robust* explanatory reasoning bias. The present paper empirically demonstrates across six experiments ($N = 2200$) that this bias is less robust than has been claimed, and influenced by nuanced pragmatic inferences on the side of participants. Pragmatic factors shown to influence the bias are assumptions about how easily an unobserved piece of evidence should have been observed if it was present ("feature diagnosability"), and the formulation of the test question being asked. Across studies, genuine narrow latent scope biases resulting from fallacious reasoning were found only in a fraction of participants. It is also demonstrated that the magnitude of the bias depends on response options: it is stronger if participants are forced to commit an error, but at best weak if they are allowed to give the correct answer.

1. Introduction

Humans have an unquenchable thirst to make sense of this world. From a young age onward we ask *why* things are as they are. Why does the sun go down (and rise again)? Why did Peter not show up for the meeting? Why does increasing the federal funds rate (usually) lower inflation? Why does this patient have this skin rash?

The ability to explain a phenomenon not only is intrinsically satisfying but also extremely practical (Gopnik, 2000; Lombrozo, 2011; Lombrozo & Vasilyeva, 2017). Explanatory reasoning enables us to learn more effectively (Williams & Lombrozo, 2010), to form generalizations and categories (Lombrozo, 2006, 2009; Waldmann, Meder, von Sydow, & Hagmayer, 2010; Williams & Lombrozo, 2010), predict the future (Lombrozo, 2011), make diagnoses (Fernbach, Darlow, & Sloman, 2011; Fernbach & Rehder, 2013; Meder & Mayrhofer, 2017; Meder, Mayrhofer, & Waldmann, 2014), and carry out interventions with which we may change the course of events in our favor (see also Lombrozo, 2012, for an overview). Although our explanatory reasoning may often only lead to fragmentary knowledge that contains gaps (see, e.g., Keil, 2006), sometimes may foster overgeneralization (Williams, Lombrozo, & Rehder, 2013b), or lead to the perception of illusory patterns (Williams, Lombrozo, & Rehder, 2011), it nonetheless equips us with extraordinary adaptive abilities.

Researchers investigating explanatory reasoning have often been studying to what extent lay people's explanations comply with what philosophers call *explanatory virtues* (see, e.g., Brewer, Chinn, & Samarapungavan, 1998; Lipton, 2004; Lombrozo, 2016). For instance, both philosophers and scientists agree that *non-circularity* and *simplicity* are hallmarks of *good* explanations. According to famous Occam's razor, to explain a phenomenon we should not postulate more theoretical entities than necessary. Another hallmark is *breadth* or *scope*. As has been noted by Paul Thagard in his book "Conceptual Revolutions" (Thagard, 1993, p. 72): "Other things being equal, we should prefer a hypothesis that explains more than alternative hypotheses. If hypothesis $H1$ explains two pieces of evidence and $H2$ explains only one, then $H1$ should be preferred to $H2$".

Psychological research suggests that these explanatory virtues are shared by lay people (Lagnado, 1994; Lombrozo, 2007; Pacer & Lombrozo, 2017; Read & Marcus-Newhall, 1993; Shimojo, Miwa, & Terai, 2020; Vrantsidis & Lombrozo, 2022), even by children as young as five years old (see, e.g., Bonawitz & Lombrozo, 2012; Corriveau & Kurkul, 2014). A typical test situation revealing a preference for explanatory simplicity (Lombrozo, 2007) is abstractly summarized in Fig. 1. As explanatory reasoning often is a form a causal reasoning (see, e.g., Lombrozo, 2010; Lombrozo & Vasilyeva, 2017; Pacer & Lombrozo, 2017; Williams, Lombrozo, & Rehder, 2013a), Fig. 1 illustrates the typical test situation using causal graphical models (Cheng & Lu, 2017; Griffiths & Tenenbaum, 2005; Pearl, 1988, 2000; Sloman, 2005; Spirtes,

E-mail address: simon.stephan@psych.uni-goettingen.de.

https://doi.org/10.1016/j.cognition.2023.105630
Received 20 May 2023; Received in revised form 26 September 2023; Accepted 27 September 2023
Available online 6 October 2023

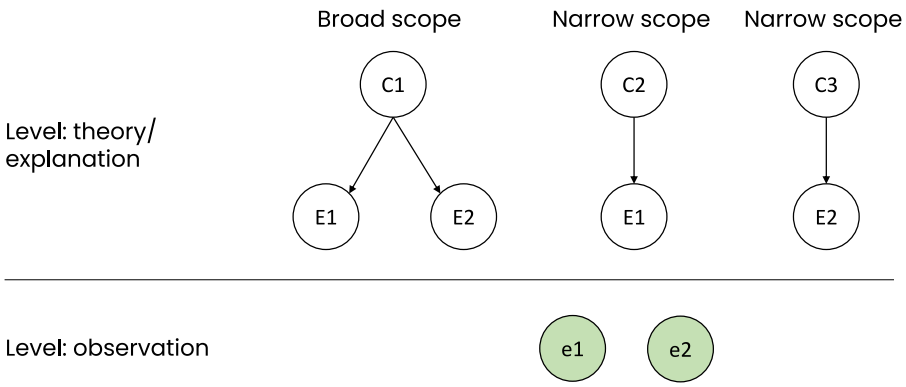

Fig. 1. Competing causal structures that explain two observed pieces of evidence (e1, e2).

Glymour, & Scheines, 1993; Waldmann, 2017), where causes and effects are depicted as nodes connected by causal arrows pointing from the causes to their effects (see also Hitchcock, 2009; Paul & Hall, 2013; Sloman, 2005). Participants observe two pieces of evidence/effects, $e1$ and $e2$ (e.g., two symptoms), which can be explained either by a single common cause ($C1$, e.g., a disease that causes both symptoms) or by a conjunct of two elemental causes ($C2$ and $C3$, e.g., two different diseases, each causing only one symptom). In such situations, it has been found that subjects tend to prefer the simple common cause explanation over the more complex conjunctive explanation (but see also Lim & Oppenheimer, 2020; Zemla, Sloman, Bechlivanidis, & Lagnado, 2017, 2023). Importantly, if the possible causes ($C1$, $C2$, and $C3$) have equal and independent base rates, such a simplicity preference is normative: given the evidence, the posterior probability of the common cause is higher than that of the conjunct of two (or more) elemental causes, $P(c1|e1, e2) > P(c2, c3|e1, e2)$. Reasoners are sensitive to these probabilities, even though they may sometimes overrely on simplicity (see, e.g., Lombrozo, 2007).

While many studies on explanatory reasoning focused on explanatory simplicity, some addressed the role of explanatory breadth (or scope) (see, e.g., Preston & Epley, 2005; Read & Marcus-Newhall, 1993). Here, too, results suggest that lay people seem to adhere to normative principles. In line with Thagard's (1993) maxim, it has been found that people seem to prefer explanations that account for a broad range of (observed/existent) phenomena over those that can account only for a narrower set of (observed/existent) phenomena. For example, when asked to explain why Cheryl has nausea, gained weight, and suffers from fatigue, subjects in a study by Read and Marcus-Newhall (1993) preferred a broad-scope explanation (pregnancy in this case) that explains all observed phenomena over a narrow-scope explanation (a stomach virus in this case) that explains only one observed phenomenon (nausea in this case).

The view that explanatory reasoning tends to be in accordance with normative principles has recently also been underpinned by a formal analysis by Wojtowicz and DeDeo (2020). In their paper, the authors argue and formally demonstrate how different phenomena of human explanatory reasoning that have been empirically observed can coherently be subsumed under the framework of normative Bayesian inference.

1.1. The narrow latent scope bias

Studies showing that lay people and philosophers tend to share the same explanatory virtues, and that lay people's explanatory preferences gravitate towards normativity (even though they might not always be optimal or perfectly rational), stand in contrast to a relatively recent series of studies on reasoners' explanatory preferences observed in situations where competing explanations differ with respect to so-called "latent scope" (Johnson, Rajeev-Kumar, & Keil, 2016; Khemlani, Sussman, & Oppenheimer, 2011; Sussman, Khemlani, & Oppenheimer, 2014). By latent scope what is meant is that some pieces of evidence predicted by a particular explanation remain unobserved in a target situation. The test situation thus differs from those studied in earlier experiments on explanatory breadth, where all phenomena of a broad-scope explanation are observed. For example, consider an explanation A that predicts two pieces of evidence, E_1 and E_2. In a particular situation, only one piece of evidence is observed (e.g., E_1) while the status of the other remains *unknown*. The unobserved piece of evidence is said to be lying in the explanation's latent scope. Studies that tested people's explanatory preferences in situations where different competing explanations vary with respect to latent scope indicate that reasoners' explanatory preferences seem to violate normative principles in these situations. These studies claim to have discovered a so-called "narrow latent scope bias" in explanatory reasoning, a non-normative preference for explanations with narrow latent scope that deviates from a probabilistic standard. This bias has been described as a "robust" (Khemlani et al., 2011; Sussman et al., 2014) reasoning bias (see Hahn & Harris, 2014, for an overview on different notions of "bias").

The narrow latent scope bias is the focus of the present paper. It pursues the goal to examine the alleged robustness of this bias. This is done by testing whether pragmatic factors influence reasoners' preference for narrow latent scope explanations, and also by looking at different kinds of test query response formats (forced choice between wrong answers vs. rating scale including the correct answer). The paper, for the first time, also aims to analyze the distribution of the bias, i.e., to assess to what extent the narrow latent scope bias is shared by the majority of people or only by a subgroup. A robust bias would imply that most subjects exhibit it and that it is at best only slightly affected by the previously mentioned factors. To foreshadow the main findings, the narrow latent scope bias is found to be influenced by pragmatic factors and to largely disappear when subjects are allowed to respond correctly. Also, it is found that a genuine latent scope bias (i.e., a preference for narrow latent scope explanations that results from a fallacious reasoning process) is only exhibited by a small fraction of reasoners, whereas most respond correctly. This latter finding adds to another recent experiment conducted by Tsukamura, Wakai, Shimojo, and Ueda (2022). As an unexpected finding, the authors also reported that only a subgroup of subjects in their experiment displayed a narrow latent scope bias. Taken together, these findings suggest that the narrow latent scope bias is less robust than past studies suggested.

An example of a situation in which the narrow latent scope bias is supposed to arise is the Amazon jungle Tokolo tribe scenario used by Sussman et al. (2014) in their Experiment 1 as one of their test scenarios:

> In the jungles of the Amazon, about half of the Tokolo tribe members are hunters, and the other half are spear fishermen. Both hunters and spear fishermen carry spears, but spear fishermen also carry nets.

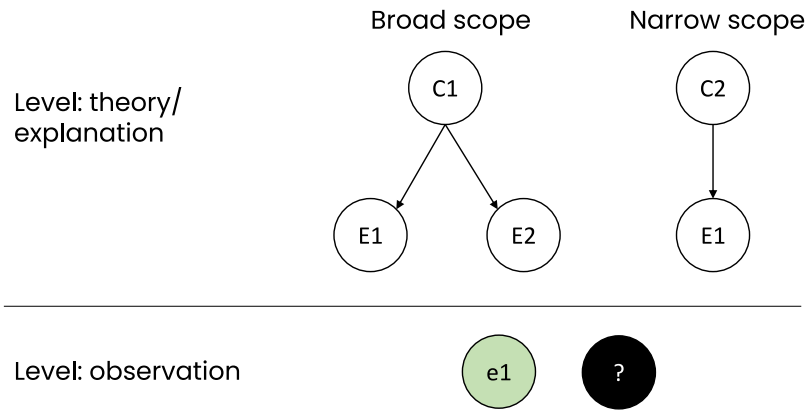

Fig. 2. Competing causal structures with different latent scope that explain one observed piece of evidence (e1, ?).

You come across a tribesman who has a spear, but you don't know whether or not he also has a net.

The question is whether the tribesman is more likely to be a hunter or a spear fisherman. Subjects had to indicate the more probable explanation by selecting one of the two possible categories. Most chose the explanation "hunter" in this case, showing what seems to be a bias for narrow-latent scope explanations.

Choosing the narrow-scope explanation is considered a non-normative bias in this case because hunters and spear fishermen are equally common in the tribe, and they each always have their respective features. The person you see, for whom the relevant/diagnostic feature remains unobserved, is objectively equally likely to be a hunter or a spear fisherman.

A formal description can be given using the abstract causal models shown in the top part of Fig. 2. They represent the two competing possible (causal) explanations for the evidence. The common cause $C1$ represents a broad-scope explanation as it predicts two effects/features (e.g., spear and net), and the single-effect cause represents a narrow-scope explanation because it predicts only a single effect/feature (e.g., spear). The unobserved effect/feature (e.g., net) is beyond the scope of the narrow-scope explanation ($C2$ in this case) but within the scope of the broad-scope explanation ($C1$ in this case). As the state of this effect/feature is unobserved, it is said to lie within the latent scope of the broad-scope explanation. The reason why a preference for one of the competing explanations is considered a bias becomes apparent if we apply Bayes' rule in the form yielding the explanations' posterior odds:

$$\frac{P(C_1|E_1)}{P(C_2|E_1)} = \frac{P(C_1)}{P(C_2)} \cdot \frac{P(E_1|C_1)}{P(E_1|C_2)} \quad (1)$$

If both causes (or explanations/hypotheses) have identical base rates, i.e., equal prior probabilities ($P[C_1] = P[C_2]$), and if the likelihood of the shared effect/feature is the same under each hypothesis ($P[E_1|C_1] = P[E_1|C_2]$), it follows that observing the shared effect/feature provides no evidence in favor of one the competing explanations; both have the same posterior probability.

$$\frac{P(C_n|E_1)}{P(C_b|E_1)} = 1 \cdot 1 = 1 \quad (2)$$

This leads to the question of how past studies have explained the narrow latent scope bias. According to Khemlani et al. (2011), "narrow latent scope explanations might be preferred because of their close match to the observed data. There are fewer predictions made by the explanation about which the observer is uncertain. This may yield a bias towards narrower scope explanations" (p. 528).

Another explanation was proposed by Johnson et al. (2016), which is called *inferred evidence*. According to the inferred evidence account, "people perform explanatory reasoning using not only the observed evidence, but also inferred evidence (Johnson, Rajeev-Kumar, & Keil, 2014). That is, when some evidence is unavailable but potentially diagnostic, people make a guess as to what that evidence would be, if it were known" (p. 43). Under this hypothesis, a narrow latent scope bias is predicted if reasoners conclude that it is more likely that the latent effect/feature is absent (see also Johnson, Johnston, Toig, & Keil, 2014), leading them to rule out the broad-scope explanation.

Importantly, according to the inferred evidence account the crucial mechanism leading reasoners to assume that the unobserved (latent) piece of evidence is more likely to be absent in the target situation is supposed to be a process of erroneous *probabilistic* reasoning. This reasoning process is assumed to be influenced by what a reasoner assumes (or knows) about the base rate of the broad-scope explanation's unobserved feature. The main idea of the inferred evidence account is that if subjects think (or know) that the unobserved effect/feature is usually absent, then they would tend to think that it is also absent in the target situation. By doing so, they would ignore the fact that one of the two possible explanations is presupposed to be true, and that both explanations can equally account for the observed effect/feature. Inferring that the latent effect/feature is more likely absent because it is generally rare, and that the narrow-scope explanation is therefore more likely to be true, means committing a statistical fallacy. As Johnson et al. (2016) have put it: "[...] this latent scope bias is qualitatively non-normative from a probabilistic standpoint". (p. 43) This can be illustrated as follows: As both C_1 and C_2 are assumed to have equal prior probabilities (which in most experimental scenarios of past studies has been explicitly stated), it can be concluded that in 50% of all cases in which the shared effect is present it is due to C_1 and in the other 50% due to C_2. Since the latent (unobserved) effect would be present in all cases in which C_1 is present (as causes are described to be deterministic), the probability that it is present in the given test case is also 50% (see also Johnson et al., 2016, p. 44). In their experiments, Johnson et al. (2016) *explicitly* manipulated the base rate of the broad-scope explanation's latent effect/feature. In line with the inferred evidence account, they found that this manipulation influenced subjects' preferences for narrow latent scope explanations.

It is questionable, however, if the observed preferences for narrow-scope explanations in other studies (e.g., Khemlani et al., 2011; Sussman et al., 2014) resulted from the faulty probabilistic reasoning process described by the inferred evidence account. The scenarios subjects read in these studies did not mention the base rate of the latent effect/feature, and it seems unlikely that subject inferred them spontaneously. For example, in the Tokolo tribe scenario introduced above, it seems implausible that subjects failed to restrict their considerations to the Tokolo tribe and its two member categories, and instead began to think about the generally low prevalence/base rate of people carrying nets.

The inferred evidence account does not seem to explain these cases, which leads to the question of what other factors make subjects prefer narrow latent scope explanations. The present paper hypothesizes that apparent narrow-latent scope biases can often be explained by participants making sensible pragmatic assumptions. It is assumed that reasoners who behave as if they were prone to a *genuine* narrow latent scope bias often have good, rational, reasons for preferring narrow latent scope explanations. By genuine narrow latent scope bias, what is meant is a preference for narrow-latent scope explanations based on fallacious reasoning. The view of the present paper thus differs from the inferred evidence account, which says that preferences for narrow latent scope explanations mostly come from a fallacious probabilistic reasoning process.

The present paper does not deny that the fallacious probabilistic reasoning process postulated by the inferred evidence is real (in fact, some of the present experiments will provide evidence for it). If subjects are given explicit information about effect/feature base rates, as in Johnson et al.'s (2016) experiments, then (at least) some subjects are (unwarrantedly) influenced by it. The main claim of the present paper is that much, but not all, of what looks to be a genuine narrow latent scope bias is due to sensible and nuanced pragmatic reasoning that pays attention to small details of the experimental scenario or/and test question.

1.2. Diagnosability of effects/features

One pragmatic factor investigated in this paper that may contribute to apparent narrow latent scope biases is the *diagnosability* of the category features. What is meant by feature diagnosability is how easily a feature's status (i.e., present or absent) can be determined. To illustrate, we may consider again the Tokolo tribe scenario. The two features of the broad-scope category are spear and net. These two objects are roughly equally easy to identify (or diagnose/observe) if present: a spear seems to be (about) as easy to spot as a fishing net. In the test situation, when we are told that we see a tribesman with a spear but we do not know if he also has a net, we might wonder how it is possible that we see the spear but not the net, assuming that we seem to be close enough to spot the spear and that a net is not harder to see than a spear.[1] One may conclude that this kind of situation is more likely to happen in a context in which the unknown feature is actually *absent*. This is because we may assume that, given that we see the spear, we should also have noted the net if it was actually present. Crucially, this seems to be a *reasonable* pragmatic conclusion rather than a fallacious/biased way of reasoning. To further illustrate the pragmatic relevance of feature diagnosability, we may contrast the original version of the Tokolo tribe scenario with one in which the diagnosability of the two features clearly differs, i.e., a scenario in which the unobserved feature is harder to diagnose than the evident feature. For example, we may imagine a scenario in which all Tokolo hunters and fishermen wear colorful feathered headdresses, and only fishermen also have a golden molar tooth. Now we may imagine a situation where we come across a tribesman and see that he is wearing a colorful feathered headdress but not if he also has a golden molar tooth. Unlike in the original scenario, it seems that we are not as perplexed by the fact that the status of the relevant feature is unclear. The reason is that, even though we spotted the colorful feathered headdress, *not* spotting the unique feature seems to be natural in this case. Importantly, not knowing the status of the unique feature in this situation seems to be just as likely to happen in a world in which this feature is present as in one in which this feature is absent. As a result, we seem to be less inclined to judge that the tribesman is probably a hunter.

This process leading to a preference for the narrow latent scope explanation via assumptions about the most likely status of the unobserved feature in the test case differs from the (fallacious statistical reasoning) process assumed by the inferred evidence account. According to the inferred evidence account, reasoners would infer that the unobserved feature is absent in the encountered tribesman not because it should be easy to see if it was present (which is a pragmatically reasonable assumption), but because it is generally (statistically) rare.

It is important to note that some past experiments (see, e.g., Experiments 2 and 3 in Johnson et al., 2016) also aimed to control for pragmatic inferences about the unobserved feature's actual status but did not find a noteworthy reduction in subjects' preferences for narrow latent scope explanations. To prevent subjects from making pragmatic inferences (e.g., about feature diagnosability), the authors gave their subjects an explanation for why the diagnostically relevant feature was unobserved. These studies will be revisited in the General Discussion, where it will be discussed as to why this manipulation might have been only weak. There, a supplementary study will be summarized, which provided evidence that giving subjects an explanation for why the relevant feature remained unobserved did reduce their preference for narrow latent scope explanations.

1.3. Test query formulation

Another pragmatic factor that may lead to apparent narrow latent scope biases is the way in which the test query is formulated. Instead of having subjects indicate which explanation of the evidence is more probable, some experiments had subjects indicate which explanation of the evidence would be more "satisfying" (see Experiment 1a in Khemlani et al., 2011; see also Experiment 1 in Johnson et al., 2016). The probability that an explanation is true may not always align with how satisfying we find it. This may be particularly acute in the experiments that used test scenarios involving aversive effects/features, such as painful symptoms or malfunctioning components of artifacts. Subjects in Experiment 1a in Khemlani et al. (2011) read a fictitious Harry Potter scenario about magic spells cast by Death Eaters that lead to different (aversive) skin alterations. Different possible spells varied with respect to the number of skin symptoms in their victims, realizing scope differences. Subjects then learned about a victim of a spell having undiagnostic symptoms, while the status of the diagnostic features was unknown. Subjects were asked to say which of the possible spells would be a more *satisfying* explanation. Subjects' tendency to pick the narrow-scope cause might have been increased in that scenario because a spell leading to fewer symptoms is better for the victim – and thus more satisfying – than a spell leading to more problems.

Similarly, Johnson et al. (2016) in Experiment 1 used a satisfaction test query in scenarios about diseases and symptoms in humans and trees, about causes of a robot's hardware problems, and about a spaceship's malfunctioning. For example, in one scenario subjects learned: "Vilosa always causes abnormal gludon levels. Pylium always causes abnormal gludon and lian levels.". Subjects then learned: "Patient #890 has abnormal levels of gludon. We don't know whether or not he has abnormal lian levels.". Subjects then rated how satisfying the different possible explanations would be. Again, one could reasonably argue that it would be more satisfying if the narrow-scope explanation was true because that would imply fewer negative symptoms.

By how much will a test question asking subjects to pick the most satisfying rather than the most probable explanation influence their explanatory choices? Two papers (Lombrozo, 2007; Vrantsidis & Lombrozo, 2022) have used and compared both types of test query formulations in experimental scenarios about diseases, where a satisfaction test query could be interpreted as a question asking for the

[1] Surprise seems to play a relevant role in these situations. We seem to begin to wonder what the status of the unobserved feature is because we are *surprised* that it is unobserved. In part, this surprise may result from the fact the diagnosability of the two features does not seem to be conditionally independent in some contexts (e.g., when we are close enough to see a spear, the probability to see a net, if it is present, should also be high).

more preferable outcome, e.g., having one rather than two diseases. These studies have found only small differences between the two kinds of test query formulations, which might mean that if there was this (mis-) interpretation of satisfaction test queries, it probably did not occur in many subjects (if many subjects did it, one would probability have observed very strong preferences for one disease over two in the satisfaction query conditions). However, these studies did not test the narrow latent scope bias but probed explanatory simplicity preferences, i.e., the causal structure of the scenario was the one shown in Fig. 1. For example, in Vrantsidis and Lombrozo's (2022) experiments, subjects read a fictitious scenario about aliens from planet Zorg who can contract three different diseases: *Tritchet's syndrome*, which causes *sore minttels* and *purple spots*, *Morad's disease*, which causes only *sore minttels*, and a *Humel infection*, which causes only purple spots. The test case described an alien suffering from both sore minttels and purple spots. Depending on condition, subjects were asked to either rate the posterior probability of each disease/explanation (or combination of diseases) or how satisfying each of the three possible explanation (or their combination) is. Importantly, unlike in narrow latent scope scenarios, possible explanations always had to account for two (present) symptoms in this scenario. As the number of symptoms that needs to be explained is the same under both the simple (common cause) and the complex (conjunct) explanation, the difference in how satisfying each of the possible explanations is might actually only be small: having only one disease (the common cause disease) may not be that much more satisfying if this one disease leads to the same number (and type) of symptoms that would also be caused by the combination of two other diseases. This is different in a narrow latent scope scenario, where one might regard the narrow latent scope explanation as more satisfying because it would imply the absence of additional aversive symptoms.

Another argument for testing the influence of test query formulation is that even a small impact of this factor might be considered relevant in cases where we "accuse" reasoners of making non-normative judgments.

1.4. Test query response format

In addition to the two pragmatic factors introduced above, which are the main focus of this paper, the present paper also looks at the role of *test question response format*. This is relevant because some past experiments (see Sussman et al., 2014) used a forced choice response format, whereas others used a rating scale format (see Johnson et al., 2016). For example, subjects in the experiments of Sussman et al. (2014) were forced to choose either the broad-scope or the narrow-scope explanation, while giving the normatively correct answer (that both explanations are equally likely) was not an option for them (see also Experiment 1d in Khemlani et al., 2011). One possibility addressed in the present paper is that test question response format might moderate the narrow latent scope bias: A forced-choice response format preventing subjects from choosing the correct answer might lead to a particularly strong narrow latent scope preference. If a strong narrow latent scope preference requires forcing reasoners to commit to a false response and disappears when reasoners are given the chance to respond correctly, the bias may be regarded as less robust than previously claimed.

Indirect evidence for this comes from experiments where subjects could respond on a rating scale that included the correct answer (see, e.g., Johnson et al., 2016). Although these experiments did not use the same experimental scenarios that Sussman et al. (2014) or Khemlani et al. (2011, see their Experiment 1d) used, comparing the results of these studies suggests that reasoners' tendency to show a narrow-latent scope bias may indeed substantially decrease if they have the chance to give the correct answer.

Why should a forced choice response format increase narrow latent scope preferences? One reason for why narrow latent scope biases might be more pronounced if subjects are forced to make a (seemingly wrong) choice is that they might be particularly motivated in this case to find reasons why one of the competing explanations must be correct. Crucially, if a test scenario (unintendedly) allows them to find plausible (e.g., pragmatic) reasons for why the unobserved effect/feature is probably absent, resulting narrow latent scope preferences might actually not be irrational. This question will be addressed in the present paper by comparing the results of Experiment 1a and b.

2. Overview of experiments

The present paper presents a series of experiments that test the impact that the factors described above have on reasoners' tendency to express a robust preference for narrow latent scope explanations. Experiments 1a-c looked at the role of feature diagnosability and, by comparing the results of Experiment 1a and b, also at the impact of test query response format. Experiments 2a-b looked at the influence of test query formulation (probability vs. satisfaction). Experiment 3 tested the narrow latent scope bias when these pragmatic factors were blocked. Also, throughout all studies, it is assessed how much evidence there is for the inferred evidence account, by looking at how many people reasoned in line with it. Experiment 3 also directly tested the inferred evidence account by comparing a condition in which subjects received explicit information about a low base rate of the relevant (unobserved) feature with a condition in which they did not. Another goal of Experiment 3 was to replicate earlier findings documenting a rational preference for explanatory simplicity.

All experimental materials, data, and analysis scripts have been made publicly available. They can be accessed via a GitHub page at https://simonstephan31.github.io/revisit_nlsbias. This website also contains demo versions of all main, supplementary, and pilot studies. All experiments were implemented as online experiments using the *jsPsych* library (de Leeuw, Gilbert, & Luchterhandt, 2023).

3. Experiment 1a

Experiment 1a, using the Tokolo tribe scenario from Sussman et al. (2014), aimed to collect initial evidence for the feature diagnosability hypothesis by comparing the original scenario to one where the latent (unobserved) category feature is clearly harder to see than the evident feature. If feature diagnosability is a pragmatic factor subjects take into account, preferences for narrow latent cope explanations should be attenuated if the unobserved feature is harder to observe. This study probed the narrow latent scope bias using the original forced choice format, however. In this case a relatively strong bias is expected, which is assumed to occur because subjects might be particularly motivated to find a (plausible) reason why their choice is correct. If this leads to a ceiling effect, a specific effect of feature diagnosability would not be visible under a forced choice test question format. For this reason, in addition to the main test query asking subjects to select one of the two explanations, subjects in this experiment were also asked to write short explanations of their forced choices, and also to indicate what they would have preferred to say if they had had the opportunity to freely answer the test question. A demo version of this experiment can be run at https://simonstephan31.github.io/revisit_nlsbias/exp1a_mat.html.

3.1. Methods

3.1.1. Participants and sample size rationale

One hundred and sixty subjects (M_{age} = 39.10, SD_{age} = 12.99, age range 20 to 72 years) recruited via the online platform www.prolific.co

participated in this online study and provided complete data. The inclusion criteria were a minimum age of 18 years, English as first language, and an approval rate (concerning subjects' participation in online studies hosted via prolific) of 90 percent. To ensure that all participants were able to understand the written instructions, prolific workers with "no formal qualifications" for the criterion "highest education level completed" were excluded from participation. Subjects also were asked to take part via PC or Laptop, and not via Tablet or Smartphone.

A sample size calculation for a one-sided one-sample binomial test against chance (chance level = 0.50) that assumes a true proportion of 0.70 in favor of the narrow latent scope explanation was carried out with the website http://powerandsamplesize.com/. This analysis showed that a sample size of $n = 37$ is required to achieve 80% test power for discerning that proportion from chance level (0.5). This number was rounded up to a target sample size of $n = 40$. Also, with $n = 40$ per condition, the proportion difference between the two experimental conditions that can be detected with 80% power is about 0.25 (assuming proportions of 0.80 and 0.55 in the two conditions; ≈0.70 across conditions). Screenshots and brief explanations of this sample size planning is provided at the repository site. All subjects gave informed consent.

3.1.2. Design, materials, and procedure

The study had a between-subjects design with two theoretically relevant conditions (an additional counterbalancing factor will be described in more detail below). The experimental factor was *feature diagnosability* (both similar vs. latent feature harder), i.e., the ease with which the two target features (i.e., the evident and the unobserved feature of the test case) can be observed. In the original scenario condition, in which the same description was presented that Sussman et al. (2014) used, the two features of the broad-scope category were *spear* and *fishing net*. Spears and nets seem to be roughly equally easily diagnosable. In this condition, the feature of the narrow-scope category that served as the unobserved (latent) feature in the test phase was fishing net. In the second condition, the two novel features that were used were the ones from the earlier example, *colorful headdress* and *golden molar tooth*. These features were assumed to differ in their diagnosability. Specifically, it was assumed that a golden molar tooth is harder to see than a colorful feathered headdress. The golden molar tooth was the feature that served as the unobserved (latent) feature in the test phase.

Subjects were alternately assigned to the different conditions. After a screen showing general study information and asking for subjects' informed consent, subjects had to confirm that they were willing to take the study seriously and that they took part via Desktop PC or Laptop. Subjects then proceeded to a new screen on which the scenario was presented. The scenario description was:

> *Please read the following (fictitious) scenario thoroughly and then answer the question below*:
>
> In the jungles of the Amazon half of the Tokolo tribe members are hunters, and the other half are spear fishermen. Both hunters and spear fishermen carry spears [wear colorful feathered headdresses], but spear fishermen also carry nets [have one golden molar tooth].
>
> You come across a tribesman who has a spear [wears a colorful feathered headdress], but you don't know whether or not he also has a net [has a golden molar tooth].[2]
>
> To which category of Tokolo tribe member do you think this tribesman more likely belongs?

Subjects had to select one of the two possible categories, *A hunter* vs. *A spear fisherman*, which were presented next to each other. The order in which the options were presented was counterbalanced between subjects. On a subsequent screen, subjects were asked to write what they would have answered if they had been free to write an answer instead of responding to a forced-choice question. The question on that screen read: "If instead of a selection via mouse click, we had asked you to freely write down whether you think the tribesman was more likely a hunter or more likely a fisherman, what would you've written?"

On a new screen subjects answered two memory check questions that were presented in a multiple-choice format with four options. The first question asked for the defining features of hunters ("What are the defining features of hunters according to the scenario you've read?") and the second for the defining features of spear fishermen ("What are the defining features of fishermen according to the scenario you've read?"). The four options for each question were: "They carry a spear and a net", "They carry only a spear", "They wear a feathered headdress and have a golden molar tooth", and "They only wear a feathered headdress".

Subjects then provided demographic data, were given the opportunity to report any technical errors they might have encountered, and finished the study on a short debriefing screen.

3.2. Results and discussion

3.2.1. Subjects' forced choices

Subjects' explanation selections are shown in Fig. 3. Replicating what Sussman et al. (2014) found, most subjects showed a narrow latent scope preference, as they chose the option saying that the test case would more likely belong to the category that does not possess the unobserved feature. It can also be seen in Fig. 3 that, unexpectedly, this happened to be the case to roughly the same degree in both feature diagnosability conditions. In the replication condition using the scenario by Sussman et al. (2014) with spear and net as target features, 80% of the subjects chose the narrow-scope category. In the novel condition where the target features were colorful feathered headdress and golden molar tooth, 93% of the subjects chose the narrow-scope category. Exact binomial tests against chance (chance level = 0.50) testing the latent scope biases in each condition were significant in both (all $p < .001$). A 2-sample test for equality of proportions indicated that the latent-scope biases did not significantly differ between the two conditions, $\Delta P = 0.925 - 0.80 = 0.125$, 95% CI $[-0.02, 0.27]$, $p_{two-sided} = .105$. Contrary to what was expected, feature diagnosability did not seem to have an effect, and subjects' forced selections did not provide evidence for an influence of feature diagnosability.

As has been speculated earlier, the absence of an effect could be due to the forced choice response format, because a forced choice response format might motivate subjects to look for additional reasons for why the narrow latent scope explanation is correct. For this reason, subjects open-ended responses were analyzed to see if there was any evidence for the psychological relevance of feature diagnosability, and also if subjects provided other plausible reasons for why the narrow latent scope explanation is correct. Finding such evidence in subjects' explanations would suggest that feature diagnosability might have an impact in experimental settings where subjects are allowed to give the correct response.

3.2.2. Subjects' explanations

The full list of subjects' explanations is included in the data files provided at the repository site. Table 1 summarizes relevant explanation

[2] In an additional, exploratory, between-subjects condition whose results are reported on the repository site, the second part of the sentence beginning with "you don't know [...]" was left out. The status of the unobserved feature thus was not mentioned at all in this condition. This condition tested another type of pragmatic inference, as the hypothesis was that not mentioning the feature would lead subjects to conclude that it is absent.

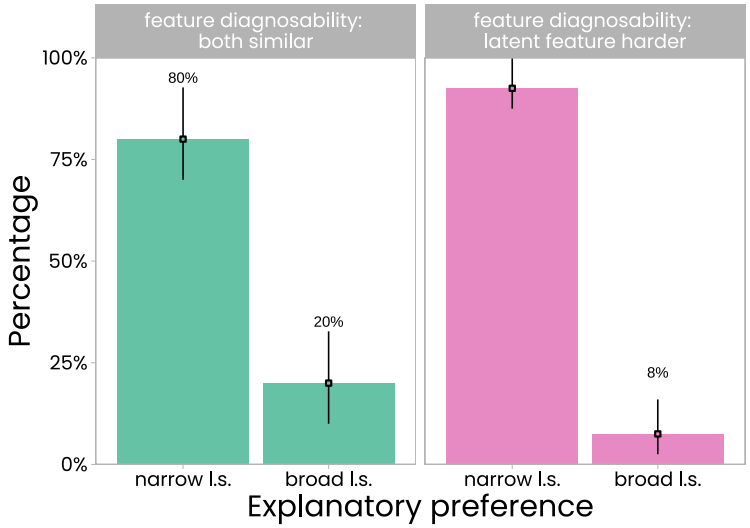

Fig. 3. Subjects' categorization ratings in Experiment 1a.
Note. Figure panels represent the two feature diagnosability conditions. Bars represent proportions of subjects having certain explanatory preferences as indicated by their ratings. Error bars represent 95% CIs of the proportions, which were computed in R using the "MultinomCI" function from the "DescTools" package with the default estimation option "sisonglaz".

Table 1
Relevant subject explanation categories identified in Exp. 1a.

Explanation category	n (%) in "Spear and net"	n (%) in "Feathers tooth"	Result of proportion test
Clearly stating that both explanations are equally likely	13 (32%)	16 (40%)	$p = .243$ (one-sided)
Clearly stating that the unobserved feature is more likely absent due to feature diagnosability	13 (32%)	2 (5%)	$p < .001$ (one-sided)
Other reason (clearly unrelated to feature visibility) for why test case more likely belongs to the narrow-scope category	1 (2%)	6 (15%)	$p = .048$ (two-sided)

categories that were identified and also shows whether these differed between the two feature diagnosability conditions. One relevant aspect of the analysis was to see to what extent these explanations indicate that subjects actually knew that the correct answer was 50:50. The explanations of a number of subjects could clearly be identified as explanations indicating that both options are normatively equally likely. This was true in both feature conditions. For example, one subject wrote: "I would have said 'don't know'. If forced to choose I would have said the tribesman was more likely a hunter on the basis that I couldn't actually see a net but could see a spear". And another wrote: "Every Tokolo tribe member carries a spear, whether or not they are a hunter or a fisherman. As 50% are hunters & 50% are fisherman there is only a 50/50 chance of the tribesman being a hunter". And yet another wrote: "It would be impossible to tell if the tribesman was a hunter or fisherman". In total, in the spear-net condition, 13 explanations (32%) clearly indicated that the test case had equal chances to belong to the broad-scope or the narrow-scope category. This number was slightly higher in the headdress-tooth condition, where 16 subjects (40%) clearly explained that both options were equally likely. A 2-sample test for equality of proportions showed that this difference was not significant, however, $\chi^2(1) = 0.49$, $p = .243$ (two-sided).

Among those subjects who explained that they considered it more likely that the test case belonged to the narrow-scope category, a number of participants wrote that they actually believed the unobserved feature to be absent. This finding was interesting with respect to the feature diagnosability hypothesis. Examples of explanations belonging to this category are: "I would have written that he was most likely a hunter as his net wasn't obviously visible", "Because I have come across him he must be in front of me and if I can't see an obvious net I am going to assume he doesn't have one and therefore is a hunter", and "A net should be visible, and while a hunter carries the right equipment to also be a spearfisherman, the spearfisherman needs a net to differentiate him/herself from the hunter". According to the feature diagnosability hypothesis, the tendency to assume the absence of the unobserved feature should overall be stronger in the spear-net scenario than in the headdress-tooth scenario, where the unobserved feature (net) should be as easily visible as the manifest feature (spear). Indeed, 13 explanations (32%) clearly stated that the unobserved feature was probably absent in the spear-net condition. By contrast, the same was true only for two explanations (5%) in the headdress-tooth condition, where not seeing the unobserved feature (a person's tooth) was assumed to be less surprising. A directed proportion test (of the same type as the one mentioned before) confirmed that these proportions significantly differed, $\chi^2(1) = 9.92$, $p < .001$ (one-sided).

There were also subjects who provided other reasons (e.g., based on plausible background assumptions) for why they considered the test case to be more likely the member of one category than the other. This might explain why no difference in subjects' narrow latent scope preferences was observed between conditions. Example explanations are: "I would say a hunter, because I am meeting them on land. There is a slightly bigger chance that the hunter is on land than the fisherman, who will sometimes be at sea, fishing", "I would have chosen hunter because a spear fisherman would likely be near water which it didn't mention", "I have no idea, and think it's a 50/50 choice. However if you suppose fishermen spend a certain amount of time in the sea, then I am more likely to bump into a hunter – therefore I chose hunter", and "I think either option is equally likely but if I have to choose one or the other then I choose hunter because I think I am more likely to encounter them because fishermen would be out on the river or sea so I am less likely to meet them perhaps". The explanations of seven subjects (six in the feather-tooth and one in the spear-net condition, $\chi^2(1) = 3.91$, $p < .048$, two-sided) could clearly be classified as belonging to this category. These explanations indicate that preventing subjects from responding with the normatively correct answer may indeed encourage them to look for (other) reasons why one of the presented (wrong) options may be correct. Taking feature diagnosability as a reason was

less of an option for subjects in the novel condition (where it is less surprising that the unobserved feature remains unobserved because it is less visible). Indeed, explanations like the ones presented above were found mostly in the novel feather-tooth condition, which indicates that subjects here began to look for (and found) alternative plausible (pragmatic) reasons in favor of the narrow-scope explanation.

A number of subjects also wrote explanations that were unclear or were no explanations. For example, some wrote "I think the tribesman was a hunter", "More likely a hunter", or "i would lean towards hunter". What drove the responses of these subjects remains unclear.

A final insight gained from subjects' explanations is that they did not provide evidence for the process assumed by the inferred evidence account: Subjects did not tend to say that they believed the unobserved feature to be absent because it is generally rare.

3.2.3. Conclusion

This experiment found pronounced latent-scope preferences under a forced-choice format. This was the case in all conditions and, contrary to what was expected, an influence of pragmatic reasoning about feature diagnosability did not yield an observable difference in subjects' explanation selections. Thus, if feature diagnosability has the potential to influence the narrow latent scope bias, it did not show in the context of a forced choice format and this particular test scenario. That said, there was initial evidence in subjects' open-ended responses suggesting that feature diagnosability was something they considered. More subjects mentioned that they expected to see the unobserved feature in the condition in which it should actually be easily visible (spear vs. net). One reason why this did not lead to an observable difference in subjects' selections is that subjects in the other condition (colorful feathered headdress vs. golden molar tooth) found other, less obvious, pragmatic reasons why the narrow latent scope explanation is correct (e.g., that it is more likely that a tribesman is a hunter if we encounter them on land). The fact that subjects in this condition found other pragmatic reasons might have overshadowed the potential effect of feature diagnosability. These qualitative findings might still be regarded as initial evidence for feature diagnosability hypothesis. Also, subjects' explanations provided evidence that they actually knew that both competing explanations are equally likely, which is why narrow latent scope biases can be expected to be smaller as soon as subjects are allowed to provide the correct answer.

4. Experiment 1b

The goal of Experiment 1b was to probe again the influence of pragmatic reasoning about feature diagnosability. The same feature pairs as in Experiment 1a were used, but this time the study used a rating scale that included the correct answer. In this case, it is predicted that narrow latent scope biases will be weak overall. It is also predicted that (at least some) subjects who are presented with a scenario in which the unobserved feature (a net) should be easily visible might still (plausibly) infer its absence, which should still lead to narrow latent scope preferences. By contrast, in a scenario in which it is not surprising that the unobserved feature (a golden molar tooth) is unobserved because it is difficult to see, subjects should be less inclined to infer its absence and, therefore, be more likely to say that both explanations are equally likely. Hence, according to the feature diagnosability hypothesis, subjects' tendency to infer the absence of the unobserved feature should be stronger in the spear-net than in the headdress-tooth condition. Thus, narrow latent-scope preferences were predicted to be greater in the spear-net than in the headdress-tooth condition.

Another goal of this experiment was to assess the distribution of the narrow latent-scope bias. This experiment therefore also addresses the question of whether the narrow latent scope bias is shown by all subjects or whether there are different subgroups. A demo version of this study can be run at https://simonstephan31.github.io/revisit_nlsbias/exp1b_mat.html.

4.1. Methods

4.1.1. Participants and sample size rationale

Two hundred subjects (M_{age} = 38.40, SD_{age} = 12.83, age range 19 to 76 years) recruited via the online platform www.prolific.co participated in this online study and provided complete data. The inclusion and exclusion criteria were the same as in Experiment 1a. Prolific workers who served as subjects in Experiment 1a were not allowed to participate.

The rationale behind the sample size was to have at least n = 50 complete data points (subject responses) per condition and to achieve a certain degree of estimation precision of the latent-scope bias. It was decided that none of the 95% CIs of the group means of each relevant condition should be wider than 1.5 points of the eleven-point rating scale. An analysis of the CI widths after the collection of 50 subjects per condition revealed that this criterion had already been reached (largest CI-width 1.19). Data collection could already be terminated at this point.[3]

4.1.2. Design, materials, and procedure

The main experimental factor was feature diagnosability, which was manipulated between subjects. As experimental scenario, the two versions of the Tokolo tribe scenario from Experiment 1a were used. Subjects were alternately assigned to the different scenario conditions.[4] An additional counterbalancing factor will be described below.

The procedure was largely identical with that of Experiment 1a. In the test phase, instead of having to choose either the narrow- or the broad-scope category, subjects were asked to provide their judgment on an eleven-point rating scale. The labels of the scale's endpoints were the two categories (hunter vs. spear fisherman). Whether the narrow-scope or the broad-scope category was displayed on the right or left endpoint of the scale was counterbalanced between subjects. The midpoint of the scale was labeled *50:50 (both equally likely)*.

Like in Experiment 1a, the test case description read: "You come across a tribesman who has a spear [wears a colorful feathered headdress], but you don't know whether or not he also has a net [a golden molar tooth]".

Subjects also provided brief explanations of their judgment on a separate screen. They also answered two memory check questions about the categories' defining features. They then provided demographic data, could report potential technical problems, and finished the study on a short debriefing screen.

4.2. Results and discussion

4.2.1. Subjects' categorization ratings

Subjects' categorization ratings are shown in Fig. 4a. Negative values indicate a preference for the narrow latent scope category. The

[3] Note that the stopping rule was based purely on the CI widths, not on means (or mean differences) or the location of the CI. Planning for estimation precision (see also Cumming, 2013, who uses the term "precision for planning") in this way precludes inflation of the type-one error rate even if data collection were to be continued after inspections or "checks".

[4] The experiment had the same additional between-subjects condition as Experiment 1a, whose results are reported on the repository site. This is why the overall sample size of this experiment was N = 200. This additional condition, where in the test case any information about the unobserved feature was omitted, was intended to provide further evidence for the influence of pragmatic reasoning. The prediction was that subjects would be even more inclined to infer the absence of the unobserved feature if it is not mentioned in the test case description. Furthermore, the influence of this factor was predicted to depend on feature diagnosability. Its impact was assumed to be weaker in the headdress-tooth condition, where the unobserved feature (golden tooth) is readily expected to remain unobserved. In this condition, not mentioning it at all in the test case description was expected to have less impact than in the spear-net scenario. This was indeed found.

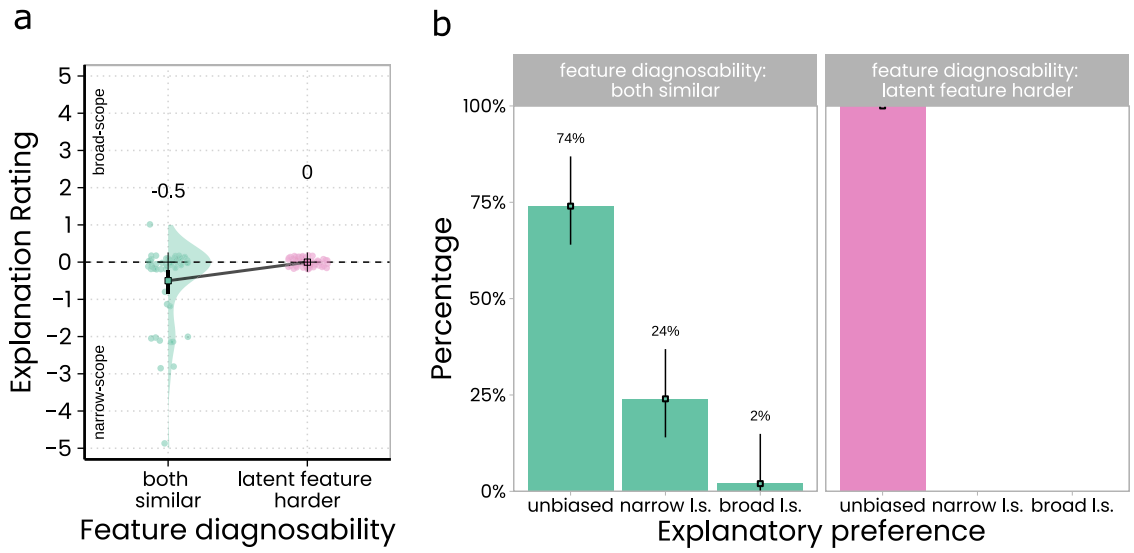

Fig. 4. Subjects' categorization ratings in Experiment 1b.
Note. a: Squares and annotations denote means, "+" denote medians. Error bars represent 95% CIs of the means. Jittered dots show subjects' individual ratings and density plots their distribution. b: Figure panels represent the two different feature diagnosability conditions. Bars show proportions of subjects having certain explanatory preferences as indicated by their ratings. Error bars represent 95% CIs of the proportions, which were computed in R using the "MultinomCI" function from the "DescTools" package with the default estimation option "sisonglaz". (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

first (green/left) plot in Fig. 4a shows the results in the condition in which the original feature pair was used (spear and net). As can be seen there, having the option to express the normatively correct answer led only to a weak latent scope bias ($M = 0.50$, 95% CI [−0.82, −0.22], Median = 0.0). As for the distribution of the latent scope bias, the jittered dots and density plots show that the biased group mean actually resulted from a minority of subjects. Most subjects did not show a narrow latent scope bias. This can also be seen in Fig. 4b, which shows the proportions of subjects with different explanatory preferences. Of the 50 participants in the spear-net condition, 37 (74%) indicated that the tribesman had equal chances to be a hunter or a spear fisherman. Twelve subjects (24%) indicated that it is more likely that the tribesman is a hunter (latent-scope preference), and one participant (2%) said that the tribesman is more likely to be a spear fisherman. As predicted by the feature diagnosability hypothesis, in the feather-tooth condition (pink/right plot in Fig. 4a), in which the unobserved feature (golden molar tooth) is harder to see than the evident feature (colorful feathered headdress) and hence more readily expected to be unobserved, subjects' narrow latent scope preferences were reduced further. In fact, they completely disappeared in this condition; all subjects indicated that both possibilities are equally likely (see also Fig. 4b). The results of a directed Welch unequal variances t-test testing the difference in subjects' mean ratings was significant, $t(49) = -3.24$,[5] $p = .001$ (one-sided). Similarly, a directed 2-sample test for equality of proportions conducted with R's *prop.test* function from the *stats* package confirmed that the proportion of subjects who responded correctly in the feather-tooth condition (100%) was significantly higher than in the spear-net condition (74%), $\chi^2(1) = 14.94$, $p < .001$ (one-sided). These results corroborate the hypothesis that in cases where category features should be equally easily observable (as was the case in the original spear-net condition), reasoners tend to take the information that the unobserved feature's status is unknown as evidence for its absence. This leads them to conclude that the test case probably belongs to the narrow-scope category.

Looking at and comparing the behavior observed in Experiments 1a and b, it seems that test question response format moderates narrow latent scope preferences. To statistically test the influence of test query response format (forced choice in Exp. 1a vs. continuous scale with correct answer in Exp. 1b), the proportions of the different behaviors were compared cross-experimentally (Figs. 4b vs. 3b). Directed 2-sample tests for equality of proportions were conducted that tested the proportions of subjects who preferred the narrow latent scope explanation in the different conditions of Experiments 1a and b, respectively. As for the original spear-net scenario condition, the proportion of subjects showing a narrow latent scope preference was significantly lower when subjects provided their judgment on a rating scale than when they were forced to choose one of the competing explanations (24% in Exp. 1b vs. 80% in Exp. 1a), $\chi^2(1) = 27.90$, $p < .001$ (one-sided). The same was true for the novel feather-tooth condition (92% vs. 0%), $\chi^2(1) = 78.54$, $p < .001$ (one-sided). These results show that preferences for narrow latent scope explanations substantially reduce under a continuous as opposed to a force-choice test query format.

4.2.2. Subjects' explanations

An analysis of subjects' explanations also corroborated the hypotheses. Table 2 summarizes relevant explanation categories that were identified. Whether these differed between the two feature diagnosability conditions is also shown. First of all, most subjects in all conditions who gave the normatively correct rating also provided accurate explanations. Three example explanations are: "If you can't establish if the person has a net then he could be either a fisherman or a hunter. The only way to determine is to see if they have a net which you do not know at this point", "I think it is a fifty fifty chance because the only information you have is that you see them with a spear. There is not enough information to decide either way", "Both groups wear colourful headwear but as I was not certain if he had the golden tooth I couldn't be sure which group he belonged to", and "100% of the tribe wear colourfull head-dresses, but the 50% who are fishers can be identified by a gold molar tooth. As we only know that this individual has a colourfull head-dress it is 50-50 whether they are a hunter or fisher". Table 2 shows that the proportion of subjects whose explanations clearly stated that the test case had equal chances of belonging to either category was higher in the feather-tooth condition. A directed 2-sample test for equality of proportions was significant, $\chi^2(1) = 11.98$, $p < .001$ (one-sided).

[5] Note that the df in a Welch unequal variances t-test deviate from those in a Student t-test.

Table 2
Relevant subject explanation categories identified in Exp. 1b.

Explanation category	n (%) in "Spear and net"	n (%) in "Feathers tooth"	Result of proportion test
Clearly stating that both explanations are equally likely	35 (70%)	48 (96%)	$p < .001$ (one-sided)
Clearly stating that the unobserved feature is more likely absent due to feature diagnosability	7 (14%)	0 (0%)	$p = .003$ (one-sided)

Among those subjects who considered it more likely that the test case belonged to the narrow-scope category, their explanations tended to indicate that the unknown feature was probably absent. Importantly, the explanations also provided evidence for the feature diagnosability hypothesis. Three example explanations of this kind are: "I thought if he was a fisherman, his net would be more visible. Given that I wasn't sure, I hedged my bets a bit and thought he was slightly more likely to be a spearsman", "I think that it is more probable that they are not a fisherman because a net would be easy to see", and "I couldn't guarantee he does not have a net with him, it could be placed somewhere in his clothing or laid down on the ground so I can not presume he is definitely a hunter. However, because he seems to appear without a net, I can try to guess that he is more likely to be a hunter than a fisherman". In the feather-tooth condition, no explanation indicated that the unobserved feature was probably absent. There were only two subjects who did not clearly write that the test case had equal chances of belonging to either category, but they did not write that they thought that the unobserved feature (the golden tooth) was absent. They wrote: "because you cannot visably notice a back tooth on someone straight away. where as a head rest is more noticable" and "it doesn't seem there is much more ways to figure out what tribe they belong to". A directed 2-sample test for equality of proportions confirmed that the proportion of subjects who wrote that the unobserved feature is more likely absent because it should be easy to see if it was present was higher in the spear-net condition than in the feather-tooth condition, $\chi^2(1) = 7.53$, $p = .003$ (one-sided).

4.2.3. Conclusion

The latent scope bias seems to be at best weak if reasoners are allowed to express the normatively correct answer. Most subjects in both conditions did not have a preference for one of the two explanations. The few who did had good (pragmatic) reasons for doing so (and tended to report them in their explanations): Subjects in the spear-net condition whose categorization ratings expressed a narrow latent scope preference tended to indicate that their ratings were actually driven by reasonable assumptions about the unobserved feature's status. Their explanations tended to express the following line of reasoning: A feature that cannot be seen in a situation where it should be just as easy to see as the evident feature is probably absent. It is, therefore, reasonable to assign a higher probability to the narrow-scope category.

5. Experiment 1c

Experiment 1b tested only a single scenario. The goal of Experiment 1c was to generalize the findings of Experiment 1b by manipulating feature diagnosability in further scenarios. The experiment tested three novel scenarios: *desert plant*, *stamp*, and *Swiss watch*. As in Experiment 1b, each scenario had two versions. In one, both category features had about equal diagnosability. In the other scenario, the unobserved feature was harder to diagnose than the observed feature.

The desert plant scenario was about a newly discovered species of desert plant (called Desert Daisy). Two sub-types were described (*Type A* and *Type B*), which were said to occur equally often. The shared feature was *pearly white blossom*. The second feature, which served as the unobserved feature in the test case, varied between the two diagnosability conditions. In the condition in which both features had similar diagnosability, this feature was *thick green leaves*. In the condition in which the unobserved feature was harder to diagnose, it was *venomous liquid in its stem*.

The stamp scenario was about a rare stamp (called Queen of the Caribbean). The feature pairs in the condition with similar diagnosability were *ultramarine* (the evident feature in the test case) and *yellow margin* (the unobserved feature in the test case). In the condition in which the unique (latent) feature was harder to determine than the shared (evident) feature, the unique feature was *yellow sticky back*.

The Swiss watch scenario was about a Swiss watch (called the Weissentanner). The shared feature in both feature diagnosability conditions was *little Swiss flag on the dial*. The unobserved features were *red digits* (similar diagnosability) vs. *red battery cover on its back* (condition where the unobserved feature was harder to diagnose).

A demo version of the experiment can be seen at https://simonstephan31.github.io/revisit_nlsbias/exp1c_mat.html.

5.1. Methods

5.1.1. Participants and sample size rationale

Two hundred and fifty-two subjects ($M_{age} = 38.85$, $SD_{age} = 13.65$, age range 18 to 74 years) recruited via the online platform www.prolific.co participated in this online study and provided complete data. The inclusion and exclusion criteria were the same as in the previous experiments. Prolific workers who served as subjects in previous experiments were not allowed to participate.

The study was planned for precision: The stopping rule for data collection was that no 95% CI of the group means for the three scenarios should be wider than 1.0. There should also be equally many subjects in each condition. This criterion was reached after $n = 42$ subjects in each condition.

5.1.2. Design, materials, and procedure

The study had two main factors, which were manipulated between subjects. One was feature diagnosability, which had two levels (both features equally easy to diagnose vs. unobserved feature harder to diagnose). The second factor was scenario, which had three levels (desert plant vs. stamp vs. Swiss watch). An additional counterbalancing factor was the orientation of the rating scale shown in the test phase (narrow-scope explanation on left side vs. right side). This led to 12 between-subjects conditions, to which subjects were alternately assigned. The experimental procedure was largely identical with that of Experiment 1b.

Subjects read one of the three different scenario descriptions, which can be found on the repository site. For example, the desert plant scenario read:

> *Please read the following (fictitious) scenario thoroughly and then answer the question below*:
>
> In the desert region of Al Amanur between China and Mongolia, botanists recently discovered a new kind of plant, the Desert Daisy. There are two subtypes of the Desert Daisy, Type A and Type B. Both occur equally often. Desert Daisy Type A has a pearly white blossom. Desert Daisy Type B also has a pearly white blossom, but unlike Desert Daisy Type A, it also has thick green leaves [it also has a venomous liquid in its stem].
>
> You're on a trip through Al Amanur desert and come across a Desert Daisy that has a pearly white blossom, but you don't know whether it also has thick green leaves [the venomous liquid in its stem] or not.

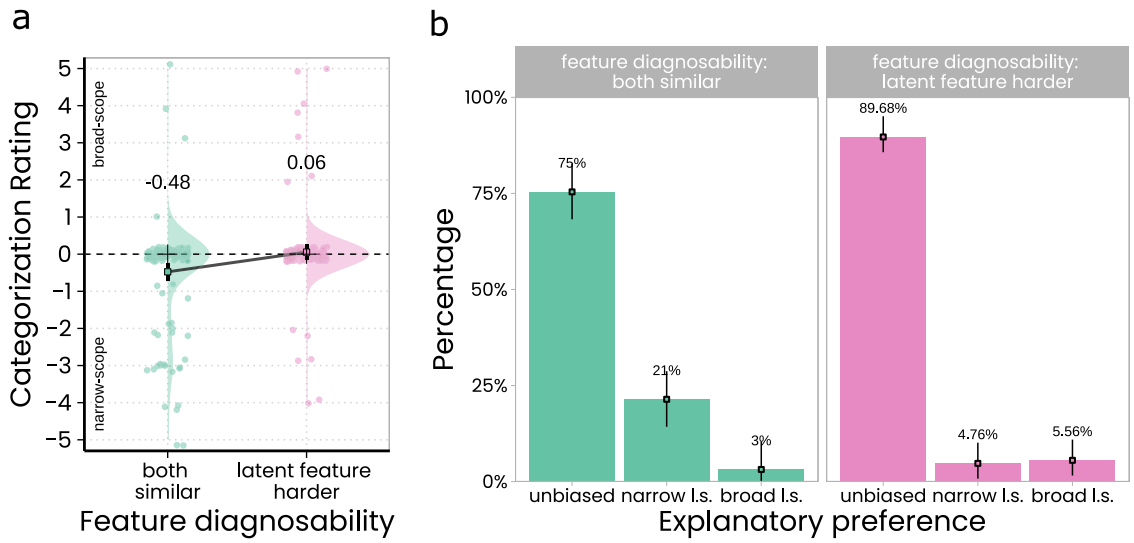

Fig. 5. Subjects' categorization ratings in Experiment 1c.
Note. a: Squares and annotations denote means, "+" denote medians. Error bars represent 95% CIs of the means. Jittered dots show subjects' individual ratings, and the density plots their distribution. b: Figure panels represent the two different feature diagnosability conditions. Bars show proportions of subjects having certain explanatory preferences as indicated by their ratings. Panels separate feature diagnosability conditions. Error bars represent 95% CIs of the proportions, which were computed in R using the "MultinomCI" function from the "DescTools" package with the default estimation option "sisonglaz".

On the same screen below the scenario description, subjects were asked to indicate on an eleven-point rating scale to which of the two possible categories they think the test object belongs to. The scale's endpoints were *Definitely Type A* and *Definitely Type B* (in counterbalanced order). The midpoint was labeled *50:50 (both equally likely)*. The remainder of the experiment was identical to Experiment 1b.

5.2. Results and discussion

The results, averaged over scenarios, are summarized in Fig. 5. Panel *a* shows subjects' categorization ratings. As can be seen there, subjects behaved like in Experiment 1b: Preferences for the narrow latent scope explanation were at best weak in all conditions (medians were 0 in all conditions), but tended to be slightly stronger when the unobserved feature should be as easy to see as the observed feature. Here, the group mean indicated a small narrow latent scope bias, $M = -0.48$, 95% CI [−0.72, −0.24]. This was not the case when not seeing the unobserved feature seemed natural, $M = 0.06$, 95% CI [−0.15, 0.26]. This pattern was found in all scenarios except the Swiss watch scenario, were almost all subjects responded normatively. The results of a directed Welch unequal variances t-test testing the observed difference between subjects' mean ratings shown in Fig. 5a was significant, $t(237.19) = -3.30, p < .001$ (one-sided), $\Delta M = -0.53$, 95% CI of ΔM [−0.85, −0.22].

Fig. 5b shows the different subgroups of participants. Like in Experiment 1b, even in the condition where both category features were similar with respect to diagnosability, narrow latent scope biases occurred only in a subgroup of participants; most participants rated that both explanations were equally likely.

5.2.1. Conclusion

Experiment 1c replicates and generalizes the findings of Experiment 1b. Subjects who are allowed to give the normatively correct answer mostly do so. The finding that most subjects who expressed a preference for the narrow latent scope explanation were in the condition in which one could reasonably assume that the unobserved feature is absent because it should be as visible as the manifest features provides further evidence for the feature diagnosability hypothesis, and for the influence of pragmatic reasoning more generally.

6. Experiment 2a

The previous studies probed the robustness of the latent scope bias by testing the influence of pragmatic reasoning related to feature diagnosability. Additionally, these studies also looked at the role of test question response format.

The goal of Experiment 2a was to test the influence of another pragmatic factor, the ambiguity of the test question asking subjects to select the *most satisfying* explanation. Crucially, whether a narrow-scope explanation is considered a satisfying explanation may deviate from its probability of being the true explanation, especially in situations in which the explanation implies something good or bad. This was the case, for example, in Experiment 1a in Khemlani et al. (2011). The experimental scenario was about magic spells that lead to aversive symptoms. Similarly, Experiment 1 in Johnson et al. (2016), was about diseases causing certain symptoms and about technical problems in artifacts causing malfunctioning parts. In such situations, it could reasonably be argued that the narrow-scope explanation is more satisfying (though not more probable) because it would imply fewer negative outcomes. For example, a patient would arguably be better off if they had a disease with fewer symptoms.

Experiment 2a revisited such scenarios and directly manipulated whether the test question asks for the most satisfying or the most probable explanation. Unlike previous studies, the present study also tested causes leading to positive outcomes, to see if that would reverse the effect. The broad-scope cause might be regarded as the more satisfying explanation in this case, as it implies more positive outcomes.

Like in the experiments by Johnson et al. (2016), the scenario descriptions also provided subjects with explicit information about the (low) base rate of the target feature (the unobserved feature in the test case). Making subjects aware of a low base rate of the target feature should increase subjects' tendency to show a narrow latent scope bias, according to the inferred evidence account. Hence, finding only a small narrow latent scope biases if subjects are asked to make probability rather than satisfaction judgments and finding that their explanatory preferences reverse if the features are positive instead of negative would suggest that the reasoning process assumed by the inferred evidence account is less dominant than has been assumed. A demo version of this study can be run at https://simonstephan31.github.io/revisit_nlsbias/exp2_mat.html.

6.1. Methods

6.1.1. Participants and sample size rationale

Three hundred and eight (M_{age} = 37.33, SD_{age} = 12.84, age range 18 to 87 years) recruited via the online platform www.prolific.co participated in this online study and provided complete data. The inclusion and exclusion criteria were the same as in the previous studies. Subjects from previous studies of this experimental series were excluded from participation.

The stopping rule for data collection was (1) obtaining a certain degree of estimation precision of group means and (2) having equally many subjects in each test query formulation condition. A pilot study (whose data and results are provided on the repository site) pretesting the materials indicated only small deviations from zero (i.e., weak biases). To detect even small biases, it was decided that no 95% CI of the mean should be wider than 1.0 points of the rating scale. For the two conditions using the probability test query, this criterion was fulfilled with n = 50 subjects in either condition (largest CI width 0.78). In the two conditions using the satisfaction test query, the required estimation precision was reached with n = 104 subjects (largest CI width 0.95).

6.1.2. Design, materials, and procedure

The study had a 2 (type of test query: most probable explanation vs. most satisfying explanation) × 2 (feature valence: negative vs. positive) between-subjects design. Subjects were alternately assigned to the conditions.

The experimental scenario was about genetic mutations altering physiological parameters. Depending on the feature valence condition, a change in these physiological parameters was either something good (an increase in life expectancy) or bad (a decrease in life expectancy). The description also emphasized that both mutations were equally likely in the population, and it also stated the prevalence (5%) of the feature that served as the unobserved (latent) feature in the test case (cf. Johnson et al., 2016). According to the inferred evidence account (Johnson et al., 2016), providing information about a low prevalence of the unobserved feature can be expected to increase preferences for narrow latent scope explanations based on fallacious probabilistic reasoning. The scenario descriptions read:

> You are a medical researcher investigating people with certain dangerous [beneficial] physiological alterations that reduce [increase] a person's life expectancy. These problems [improvements] are caused by two different gene mutations, (1) Mut-Bic2 and (2) Mut-Taw4.
>
> - Mut-Bic2 always causes abnormal [healthy] Gludon blood levels.
> - Mut-Taw4 always causes abnormal [healthy] Gludon blood levels and abnormal [healthy] Lian blood levels.
>
> A study with 200 participants found that 10 of them had abnormal [healthy] Lian blood levels (i.e., 5%). It is also known that Mut-Bic2 and Mut-Taw4 are equally likely to occur in a person. That means, the number of people in the general population who have Mut-Bic2 is the same as the number of people who have Mut-Taw4.

The labels of the fictitious physiological effects (Gludon and Lian blood levels) and the word "abnormal" were borrowed from Johnson et al. (2014). It was assumed that the word abnormal is likely to be considered something negative in the context of genetic mutations altering blood substances.

After having read the scenario description, subjects had to pass a comprehension test that probed their knowledge of (1) the mutations' effects (i.e., their differences in causal scope), (2) the prevalence of the symptom that served as the unobserved feature in the test phase, (3) the mutations' prevalence, and (4) the effects' valence. Subjects could not proceed until they answered all these questions correctly. Subjects who answered any of the questions wrong were shown the scenario description again. Only the data of subjects who needed maximally three rounds to pass the comprehension were considered complete and used for analysis.

The description of the test case was:

> *Now consider the following situation and then answer the test question below:*
>
> Patient #53 has either Mut-Bic2 (which causes abnormal [healthy] Gludon levels) or Mut-Taw4 (which causes abnormal [healthy] Gludon and abnormal [healthy] Lian levels). The patient has already been found to have abnormal [healthy] Gludon blood levels, but we don't know yet whether or not the patient also has abnormal [healthy] Lian blood levels.

Depending on condition, the test question was "Which of the two possible mutations is the most probable cause of the physiological condition of Patient #53?" (probability condition) or "Which of the two possible mutations would be the most satisfying explanation for the physiological condition of Patient #53?" (most satisfying condition). Subjects provided their ratings on an eleven-point rating scale whose endpoints were labeled *Definitely Mut-Bic2* and *Definitely Mut-Taw4*. Which label was on the left and which on the right side of the scale was counterbalanced between subjects. Depending on condition, the scale's midpoint was labeled *Both equally likely* or *Both equally satisfying*.

As in previous studies, subjects also were asked to explain their rating. As this study contained a comprehension test in the beginning, no further memory check questions were asked at the end of the study.

6.2. Results and discussion

6.2.1. Subjects' ratings

Subjects' ratings are shown in Fig. 6. The means show that subjects' ratings were overall only weakly biased. Also, both the medians and modes were 0 in all conditions. This result is in line with the previous study showing that the tendency for biased responses strongly decreases if subjects are allowed to give the correct answer. Interestingly, this result was obtained even though the scenario description contained information about a low base rate of the unobserved feature, the crucial factor that according to the inferred evidence account should lead to more pronounced narrow latent scope biases.

Fig. 6 also shows an influence of test question type. The left panel in Fig. 6 shows that the smallest deviations from zero were observed in the probability conditions, where participants were asked which explanation is more probable. As is indicated by the 95% CIs, both group means were estimated with high precision but still included zero, (M = −0.22, 95%CI [−0.48, 0.02] in the negative valence condition; M = 0.12, 95%CI [−0.26, 0.56] in the positive valence condition). Stronger deviations were observed in the satisfaction conditions, where subjects were asked to say which explanation would be more satisfying. Moreover, the group means' deviations from zero were in the predicted direction: subjects tended to prefer the narrow-scope explanation if the features were negative (M = −0.60, 95%CI [−0.93, −0.29]). By contrast, when the features were positive there was a slight tendency to select the broad-scope explanation. However, the estimation interval still comprised zero as the plausible true value in this condition (M = −0.22, 95%CI [−0.11, 0.82]). Also, contrast analyses comparing the mean differences between positive and negative valence conditions separately for each query type condition showed that the means in the probability query conditions did not significantly differ from each other (ΔM = −0.34, 95% CI [−1.08, 0.39], $t(304)$ = 0.91, p = .36), whereas they did significantly differ in the satisfaction query condition (ΔM = −0.96, 95% CI [−1.47, −0.45], $t(304)$ = 3.72, p < .001).

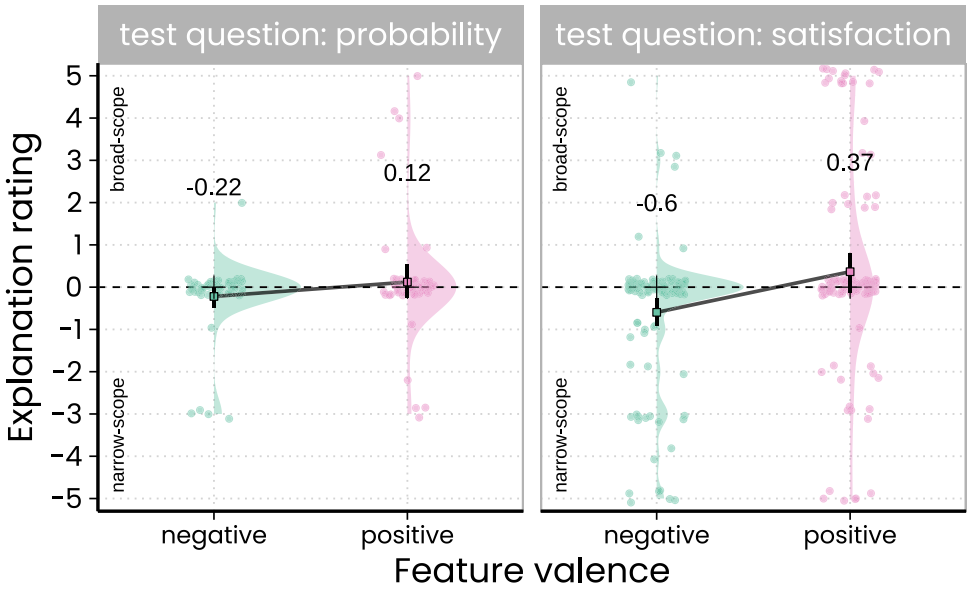

Fig. 6. Subjects' explanation ratings in Experiment 2a.
Note. Figure panels represent the two test query formulation conditions. Squares and annotations denote means, "+" denote medians. Error bars represent 95% CIs. Jittered dots show subjects' individual ratings, and density plots their distribution.

Although Fig. 6 shows that the predicted rating pattern was observed, a factorial Type III ANOVA with the two between subjects factors *test query formulation* and *feature valence* conducted in *R* with the package *afex* (Singmann, Bolker, Westfall, Aust, & Ben-Shachar, 2022) did not yield a significant interaction effect between test query formulation and feature valence, $F(1, 304) = 1.87$, $p = .172$, $\eta_p^2 = .006$. The only significant effect was a main effect of feature valence, which resulted from a preference for the broad-scope explanation when the gene mutations had positive physiological effects. This tendency occurred in for both types test question formulation, although it was more pronounced in the satisfaction condition.

Another finding revealed by Fig. 6 is that the variance in subjects' explanation ratings was generally higher in the satisfaction condition, which may suggest that the satisfaction test query was indeed more ambiguous than the probability test query.

As in previous experiments, the proportions of subjects with different explanatory preferences in the different conditions were also analyzed. Fig. 7 shows that the proportion of subjects how gave non-normative responses was higher in the satisfaction condition. Also, as predicted, when the features were negative (symptoms reducing life expectancy) the proportion of subjects preferring the narrow latent scope explanation in the satisfaction condition was higher (26%) than the proportion of subjects preferring the narrow latent scope explanation in the probability condition (10%). A directed 2-sample test for equality of proportions conducted with R's *stats* package confirmed that this proportion difference was significant, $\chi^2(1) = 4.30$, $p = 0.011$ (one-sided). When the features were positive, as predicted, the proportion of subjects preferring the broad-scope explanation was higher in the satisfaction condition (24%) than in the probability condition (12%), $\chi^2(1) = 3.04$, $p = 0.04$ (one-sided).

6.2.2. Subjects' explanations

As for subjects' explanations, Table 3 summarizes relevant explanation categories and also shows whether these differed between the two test query formulation conditions. The results of the significance tests reported in the table are the results of 2-sample tests for equality of proportions (the exact test statistics are provided in the corresponding analysis script on the repository site).

In the probability conditions, most of subjects' explanations (74%) clearly indicated that both explanations are equally likely (cf. Table 3). This suggests that most subjects who gave unbiased ratings did so based on a correct understanding of the scenario. Example explanations are: "I said both are equally likely because both alterations cause abnormal Gludon levels. If an individual has abnormal Gludon levels, it is therefore not possible to tell which of the two alterations the person could have. I was also told that both are equally likely to occur in a person", and "Given that both abnormalities occur at equal rates, it can be assumed that there is an equal chance that the patient has either one of these cell problems. The only way to find out for sure is by carrying out further tests". In the satisfaction condition, most explanations could not be assigned to this category. In part this was because subjects tended to speak about satisfaction rather than probability, sticking to the terminology they had seen in the test query.

A crucial question was whether the explanations of subjects in the satisfaction query conditions provided evidence that their ratings were influenced by pragmatic considerations about what would be a favorable result for the patient described in the test case (second row in Table 3). Clear cases of such explanations could be identified. Also, their proportion was significantly higher in the satisfaction (20%) than in the probability (2%) condition. Examples in the negative features conditions are: "As we already know that they have abnormal Gludon blood levels, it would be ideal that they simply had Mut-Bic2 as that only affects those Gludon levels. If they had Mut-Taw4 it would mean that their Gludon levels are not the only problem to worry about", "it would more more satisfying to know that he only has one, rather than both. Although either result is equally likely", and "Would rather not have abnormal Lian levels too". Example explanations in the positive features condition are: "Because both the gene mutations have gludon in them, I think it was better to have the chance of Lian it, to give the patient a better chance" and "If they have the one i selected then they would have both of the good blood things!" Some subjects in this condition even described why it would be favorable if the broad-scope cause was present, even though statistically the chances are 50:50: "because mut-taw4 always causes both healthy gludon blood levels and healthy lion blood levels so that is the best to have the mut-taw4 in your blood. But there is only a 50-50 chance that any person could have both healthy blood levels".

Explanations of subjects who preferred the narrow latent scope explanation were also screened for evidence of the reasoning process assumed by the inferred evidence account (last row in Table 3). Such

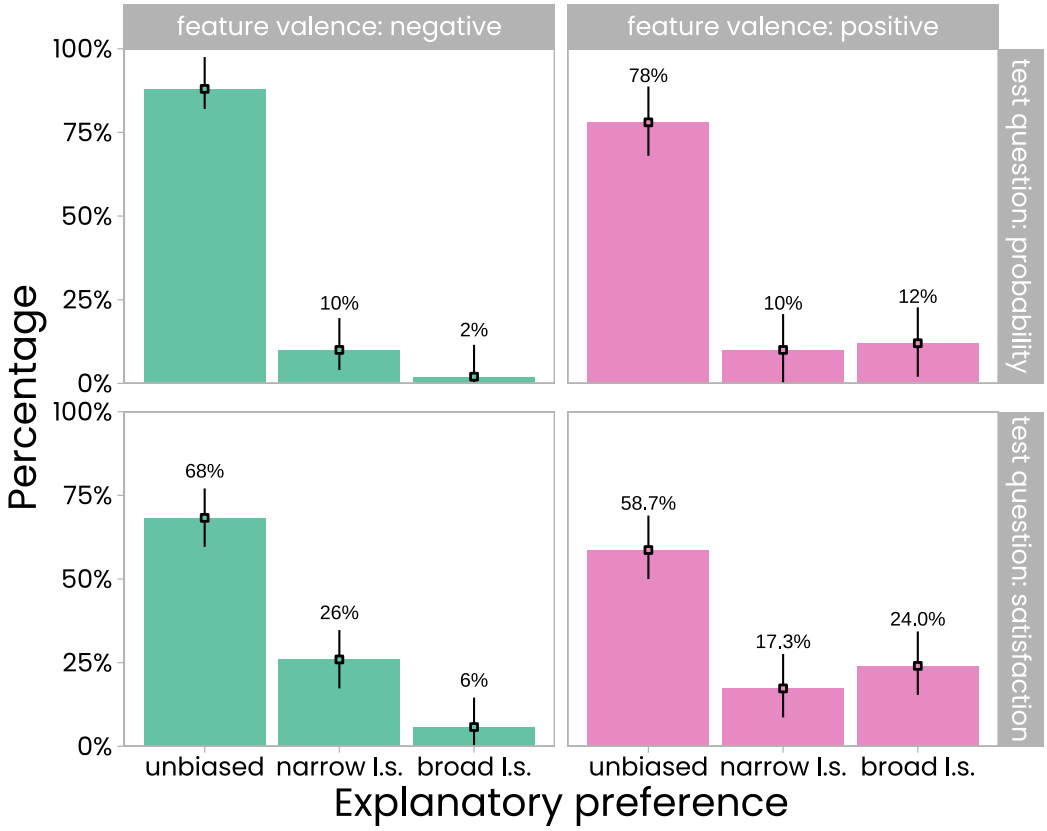

Fig. 7. Proportions of subjects having certain explanatory preferences in Experiment 2a.
Note. Figure columns represent the different feature valence conditions and rows the different test query formulation conditions. Error bars represent 95%CIs of the proportions, which were computed in R using the "MultinomCI" function from the "DescTools" package with the default estimation option "sisonglaz".

Table 3
Relevant subject explanation categories identified in Exp.~2a.

Explanation category	n (%) in "probability"	n (%) in "satisfaction"	Result of proportion test
Clearly stating that both explanations are equally *likely*	74 (74%)	97 (47%)	$p < .001$ (two-sided)
Clearly stating that the patient's perspective was the relevant aspect	2 (2%)	42 (20%)	$p < .001$ (one-sided)
Clearly describing the reasoning process assumed by inferred evidence account	0 (0%)	6 (3%)	$p = .09$ (two-sided)

explanations would express that there is a higher probability of the unobserved feature's absence in the test case because of its overall low probability. Six explanations were found that clearly fell into this category. Two examples are: "I think that the patient has elevated gludon levels so he is likely to have one of the alterations. The test for Lian hasn't been received and so is currently open ended. 5% of the population have abnormal lian levels so it is a lower chance of the patient having both gludon and lian deficiency" and "Given the previous data that only 10 of 200 people had higher Lian levels (in spite of the equal occurrence of each gene mutation), I felt that it was most likely to be this". These explanations show that the fallacious probabilistic reasoning process postulated by the inferred evidence account influenced the ratings in at least some participants.

6.2.3. Conclusion

As was the case in Experiments 1b and 1c, this study yielded narrow latent scope biases that were at best small. Most subjects in all conditions gave unbiased ratings and provided corresponding explanations. This was the case even though the scenario descriptions implied a low probability of the unobserved feature, which according to the inferred evidence account should increase narrow latent scope biases.

Also, even though the predicted interaction effect in subjects' mean ratings was not significant, the results provide tentative evidence that a satisfaction test query in combination with category features that have a valence (i.e., are negative or positive) may have prompted at least some participants to give ratings that seem biased – although they are reasonable under a satisfaction interpretation of the test query.

One potentially problematic aspect of this experiment is that the formulation of the satisfaction test query was "Which of the two possible mutations *would be* the most satisfying explanation for the physiological condition of Patient #53?" rather than "Which of the two possible mutations *is* the most satisfying explanation for the physiological condition of Patient #53?". It is possible that a "would be the most satisfying" formulation might have encourage subjects to take the patient's perspective more than a "is the most satisfying" formulation would have. In fact, previous studies used the latter formulation.[6] This problem was addressed in a follow-up study, Experiment 2b.

7. Experiment 2b

Experiment 2b addressed Experiment 2a's shortcoming by adding another test query formulation condition to the study design. In this additional condition, subjects were asked "Which of the two possible mutations *is* the most satisfying explanation for the physiological condition of Patient #53?".

[6] I would like to thank an anonymous reviewer for this observation.

A second goal of this study was to test the effect of test query formulation (satisfaction vs. probability) with higher statistical power than Experiment 2a. Also, unlike Experiment 2a, Experiment 2b only used the negative effects version of the test scenario. An influence of test query formulation would be revealed if subjects' tendency to select narrow latent scope explanations happened to be stronger in the satisfaction conditions than in the probability condition.

A demo version of the experiment is provided on the repository at https://simonstephan31.github.io/revisit_nlsbias/exp2b_mat.html.

7.1. Methods

7.1.1. Participants and sample size rationale

Seven hundred and twenty (M_{age} = 41.12, SD_{age} = 12.99, age range 18 to 78 years) recruited via the online platform www.prolific.co participated in this online study and provided complete data (n = 240 per condition). The inclusion and exclusion criteria were the same as in the previous studies. Subjects from previous studies of this experimental series were excluded from participation.

The sample size was determined in an a priori power analysis conducted with R's *pwr* package. The goal was to achieve at least 80% test power for an independent t-test testing the mean in the probability condition against the one of the novel "is most satisfying" condition. The analysis was based on the results of Experiment 2a (the mean difference observed in the negative outcome scenarios between the satisfaction and probability conditions) but assumed a slightly smaller difference: The assumed difference used in the power analysis was Δ_M = (−0.2)−(−0.55) = 0.35, and the standardizer was 1.5. This yielded an effect size of d = 0.233. The analysis revealed that the desired power is achieved with n = 228 subjects per condition. The reason why n = 240 subjects were tested in each condition was that two additional criteria were applied: (1) no 95% CI of the means should be wider than 0.5 points on the rating scale and (2) all conditions should have the same number of subjects. Both additional criteria were reached after n = 240 subjects, which terminated the data collection.

7.1.2. Design, materials, and procedure

Three test query formulations were manipulated between-subjects (*is most probable* vs. *is most satisfying* vs. *would be most satisfying*). Subjects were alternately assigned to the conditions. The scenario and procedure were identical to the one of Experiment 2a, except for the fact that only the negative effects scenario was used.

Depending on condition, subjects were asked one of the following three different test questions: "Which of the two possible mutations is the most probable explanation of the physiological condition of Patient #53?", "Which of the two possible mutations is the most satisfying explanation for the physiological condition of Patient #53?", or "Which of the two possible mutations would be the most satisfying explanation for the physiological condition of Patient #53?". Answers were provided on the same eleven-point rating scale used in Experiment 2a.

7.2. Results and discussion

7.2.1. Subjects' ratings

Subjects' explanation ratings are shown in Fig. 8. Like in the previous experiments that used a rating scale, most subjects did not prefer one explanation over the other.

Replicating Experiment 2a, the smallest amount of bias occurred in the "is most probable" condition (M = −0.15, 95% CI [−0.30, 0]). The small average bias observed here may have resulted from subjects who were indeed influenced by the information about the low base rate of the unobserved feature that was given in the scenario description, as predicted by the inferred evidence account (Johnson et al., 2016).

The jittered dots in Fig. 8a and the density plots in Fig. 8b show that, as predicted, ratings that expressed a narrow latent scope preference were slightly more frequent in the conditions in which subjects were asked about satisfaction. A Type 3 factorial ANOVA confirmed an effect of test query formulation, $F(2, 717)$ = 7.12, p < .001, η_p^2 = .19. However, Fig. 8 also shows that latent scope preferences were slightly less pronounced in the novel "is most satisfying" formulation condition (M = −0.71, 95% CI [−0.96, −0.49]) than in the "would be most satisfying" formulation condition (M = −0.71, 95% CI [−0.96, −0.49]). A Welch two sample t-test comparing the two satisfaction conditions confirmed that the mean in the novel "is most satisfying condition" was higher than in the "would be most satisfying" condition, $t(477.85)$ = 1.83, p = .034 (one-sided), d = 0.17. Yet, a second directed Welch two sample t-test comparing the novel "is most satisfying" with the "is most probable" condition showed that the novel satisfaction condition still produced a significantly larger bias than the probability condition, $t(410.33)$ = 1.86, p = .032 (one-sided), d = 0.18.

Subjects' ratings were also grouped into the three different possible response categories, which are shown in Fig. 9. The largest category in all conditions consisted of subjects how gave correct responses. The largest number of correct responses (83.3%) was observed in the "is most probable" condition. Directed equality of proportion tests confirmed that this proportion was higher than in the "is most satisfying" condition (69%), $\chi^2(1)$ = 13.30, p < .001 (one-sided) and higher than in the "would be most satisfying" condition (67%), χ^2 = 17.78, p < .001 (one-sided). At the same time, the proportions of unbiased responses in the two satisfaction conditions did not significantly differ, $\chi^2(1)$ = 0.34, p = .28 (one-sided).

The opposite result was obtained when testing the proportions of subjects who preferred the narrow latent scope explanation. This proportion was higher in the "is most satisfying" condition (22%) than in the "is most probable" condition (11%), $\chi^2(1)$ = 10.14, p < .001 (one-sided). The same was true the "would be most satisfying" condition (28%), $\chi^2(1)$ = 20.29, p < .001 (one-sided). These proportions did not differ in the two satisfaction conditions, however, $\chi^2(1)$ = 1.89, p = .17 (two-sided).

7.2.2. Subjects' explanations

Table 4 summarizes relevant explanation categories that were identified. As the two satisfaction conditions were most relevant in this experiment, the equality of proportion test results reported in the table come from tests that contrasted these two conditions. Comparing this table with Table 3, it can be seen that the results for the probability and the would be most satisfying conditions are similar to those obtained in Experiment 2a. All in all, subjects tended to say that both explanations were equally likely, especially the subjects in the probability query condition. An example for an explanation in this category is: "Both mutations are just as prevalent in the general population and both increase blood gludon levels. Therefore, without knowing the patients other blood levels, the chances of having each gene are just as likely".

Experiment 2a had revealed that a number of subjects in the satisfaction condition justified their choice by taking the perspective of the patient mentioned in the test case. Subjects who did so, and who preferred the narrow latent scope explanation, had interpreted the test query in a way that made it reasonable to select the narrow latent scope explanation. Explanations clearly falling into this category were found again in the "would be most satisfying" condition of the present study. Examples are: "I assume it's better to just have one of the mutations and not both", "Having just one issue is more satisfactory than having multiple issues", "it is better to have one abnormality than two", and "Because surely its worse to have both than just one?"

Of particular interest was whether explanations that clearly fall into this category could also be found in the novel "is more satisfying" condition. Those explanations were indeed found, even though they were rarer than in the "would be most satisfying" condition. 5% of subjects' explanations in this condition could clearly be identified as belonging to this category. Also, the proportion of these explanations was significantly smaller than in the "would be most satisfying" condition (see Table 4), but it was significantly higher than in the probability

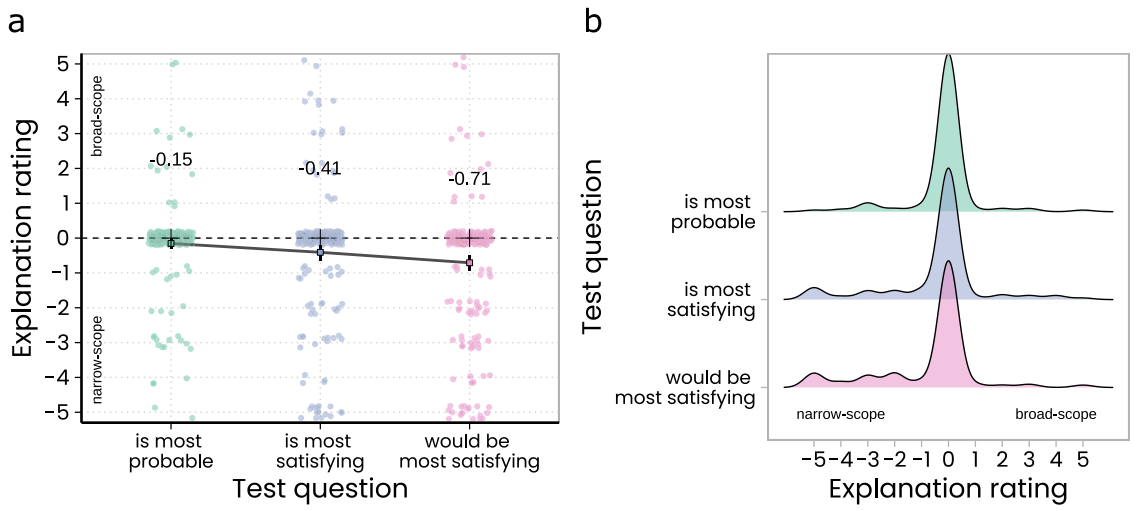

Fig. 8. Subjects' explanation ratings in Experiment 2b.
Note. a: Squares and annotations denote means, "+" denote medians. Error bars represent 95% CIs. Jittered dots show subjects' individual ratings. b: Density plots showing rating distributions.

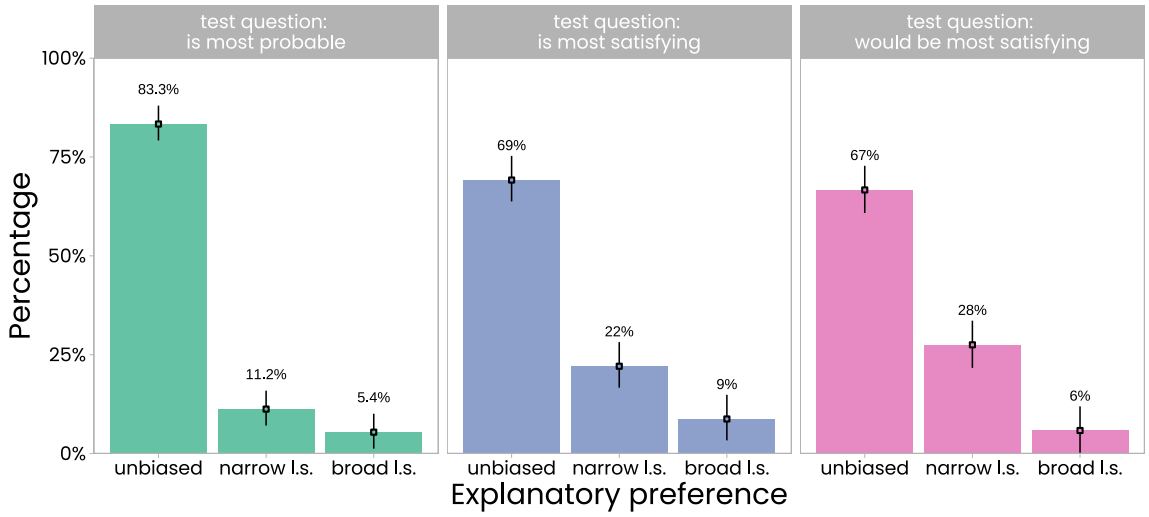

Fig. 9. Proportions of subjects having certain explanatory preferences in Experiment 2b.
Note. Figure panels represent the three test query formulation conditions. Error bars represent 95% CIs of the proportions, which were computed in R using the "MultinomCI" function from the "DescTools" package with the default estimation option "sisonglaz".

Table 4
Relevant subject explanation categories identified in Exp. 2b.

Explanation category	n (%) in "is most probable"	n (%) in "is most satisfying"	n (%) in "would be most satisfying"	Result of proportion test
Clearly stating that both explanations are equally likely	169 (70%)	139 (58%)	123 (51%)	$p < .07$ (one-sided)
Clearly stating that the patient's perspective was the relevant aspect	0 (0%)	13 (5%)	32 (13%)	$p = .001$ (one-sided)
Clearly describing the reasoning process assumed by inferred evidence account	2 (1%)	5 (2%)	4 (2%)	$p = .74$ (two-sided)

Note. Reported proportion tests compared the two satisfaction conditions.

condition, $\chi(1) = 34.29$, $p < .001$ (one-sided). Examples of such explanations are: "The question asked what was the most satisfying answer for the increased Gludon, which was the first mutation", "It would be better if the patient had this mutant, since then the patient would only have one condition" and "I think it would be better to just have the one condition than both, therefore I have said that it would be better to have this than both, although obviously better to have neither".

Also, there were explanations in both satisfaction conditions that directly mentioned the ambiguity of the test query formulation. Two examples are: "I don't have a preference for either genetic trait, and either one will provide an explanation. Each is as likely as the other. I'm also not entirely sure what 'satisfying' means in this scenario to be honest. Knowing which gene mutation is present doesn't satisfy me", "i was not sure exactly what was meant by satisfying by i rated middle as they were both equally plausable", "The word 'satisfying' is a bit ambiguous. Maybe 'plausible' would be better. As both mutations can cause the harmful Gludon levels, they are both plausible".

As in Experiment 2a, explanations of subjects who preferred the narrow latent scope explanation were also screened for evidence of the reasoning process assumed by the inferred evidence account. Explanations in line with the inferred evidence account would describe

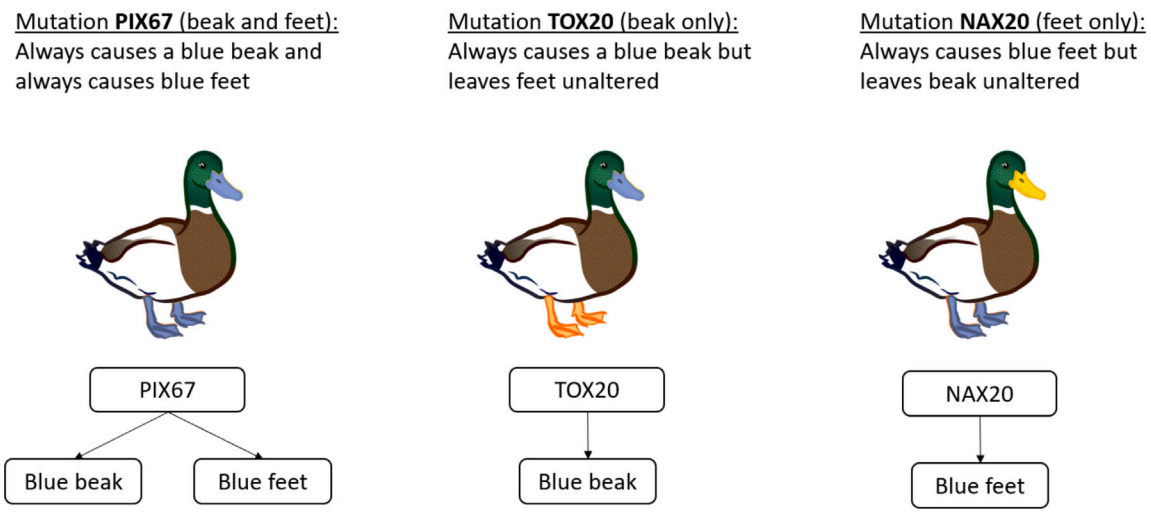

Fig. 10. Illustration of the scenario used in Experiment 3. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

higher probability of the unobserved feature's absence in the test case because of its overall low probability (low base rate of ≈5%). A few such explanations could clearly be identified. Across conditions (see Table 4), eleven subjects clearly stated that the unobserved symptom's low prevalence determined their explanatory preference. Examples are: "Only 5% of people were found to have abnormal Lian blood levels, so I thought it was more likely that the patient had Mut-Bic2 than Mut-Taw4" and "as only 10 in 200 have abnormal lian bloodlevels its more slightly likely to be mut-bic2 but you cant say for certain".

7.2.3. Conclusion

Like in previous experiments of this paper that allowed subjects to give the normatively correct answer, most subjects did. The majority of subjects who had a preference for the narrow latent scope explanation were found in the two conditions that used ambiguous satisfaction test query formulations. The "is most satisfying" formulation led to less bias than the "would be most satisfying" formulation. Although the influence of the "is most satisfying" formulation was weak in the present study, this study nonetheless shows that reasoners' behavior is sensitive to nuanced changes in test query formulations.

8. Experiment 3

Experiment 3 pursued multiple goals. As the previous experiments in this paper have demonstrated that pragmatic reasoning influences subjects' explanatory preferences, one goal was to shield subjects' responses from pragmatic factors as much as possible.

A second goal was to test again the influence of low feature rates, which according to the inferred evidence account should lead to stronger narrow latent scope biases. While Experiments 2a and 2b provided subjects with explicit information about low feature rates, this factor was not experimentally manipulated. To directly test the impact of low feature rates, Experiment 3 contrasted a condition in which subjects received information about low feature rates with one in which they did not.

A final goal of Experiment 3 was to use a scenario that not only allows testing the latent scope bias but also the replication of previous findings showing that reasoners tend to prefer simpler over more complex explanations if simpler explanations are more likely to be true than complex explanations. By demonstrating only weak latent scope biases and by simultaneously revealing preferences for simpler explanations, this experiment would provide further, converging, evidence that reasoners' explanatory preferences tend to gravitate towards normativity.

To realize a combined test of simplicity preference and narrow latent scope bias, a scenario with three alternative explanations was developed: the scenario described three independent mutations causing color alterations in ducks. An overview is shown in Fig. 10. One mutation, which instantiates a common cause structure, leads to a blue beak and blue feet. Two additional single-effect mutations either lead to a blue beak or to blue feet, respectively. To probe narrow latent scope biases, two alternative test pictures were developed, which are shown in Fig. 11. Both narrow latent scope test pictures showed a duck in the water. In one, the duck was swimming in the lake with its feet underwater. In the other, the duck was foraging on the lake with its head under water and its feet sticking out. In the scenario description, it was mentioned that these pictures were captured by a wildlife photo trap that takes pictures as soon as it detects movement. It was assumed that this information about how the evidence is collected, together with the use of non-verbal test pictures, would minimize the influence of pragmatic assumptions concerning the status of the unobserved feature.

To probe subjects' explanatory preferences in cases without unobserved features, additional test pictures were created. In these pictures, the ducks were standing at the lakeside so that their whole body was visible. Four different pictures were shown, the three displayed in Fig. 10 and one showing a duck with orange feet and a yellow beak (no mutation). The test picture probing a simplicity preference was the one showing a duck with both a blue beak and blue feet. This picture is shown in Fig. 12. A demo version of this experiment can be run at https://simonstephan31.github.io/revisit_nlsbias/exp3_mat.html. Also, a pilot study (and its results) that guided the construction of Experiment 3 is described on the repository site.

8.1. Methods

8.1.1. Participants and sample size rationale

Five hundred and sixty subjects (M_{age} = 41.10, SD_{age} = 13.92, age range 18 to 85 years) recruited via the online platform www.prolific.co participated in this study and provided complete data (n = 280 in each theoretically relevant condition). The inclusion and exclusion criteria were the same as in the previous studies.

The sample size was determined based on an a priori power analysis and the adoption of a sequential testing strategy (see Lakens, 2022, for information on sequential testing). The study was supposed to yield 80% test power for the detection of a small standardize difference of d = 0.188 between the narrow latent scope test case ratings in the two feature base rate information conditions. The effect size was obtained with the *pwr.t.test* function from the *pwr* package. The

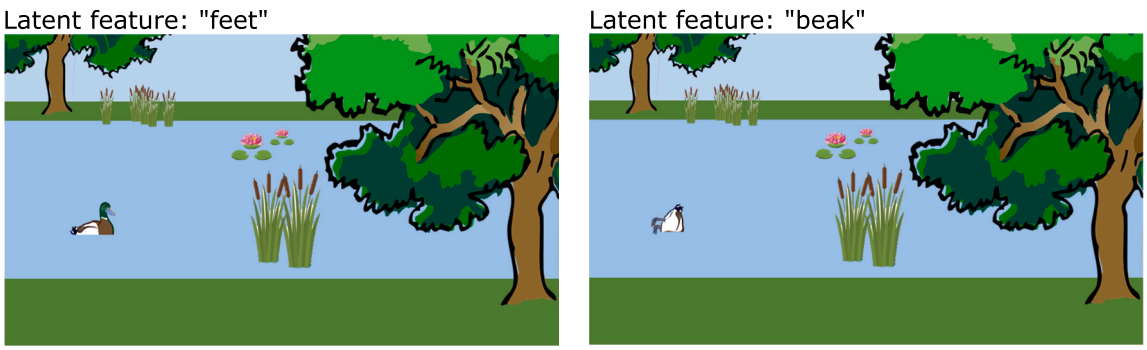

Fig. 11. Illustration of the two different latent-feature test stimuli used in Experiment 3. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

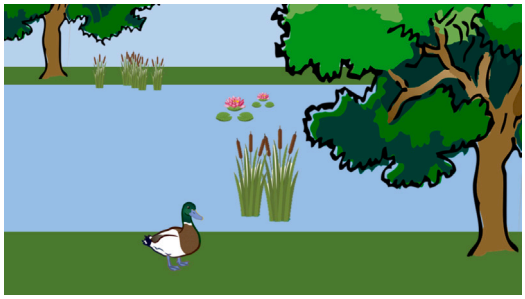

Fig. 12. The test picture in Experiment 3 that probed subjects' preference for explanatory simplicity. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

inserted expected means were −0.2 (for a small deviation from zero in the condition without feature rate information) and −0.5 (for a slightly larger deviation from zero in the condition with feature rate information). The standardizer that was used was $SD = 1.6$ (a value informed by previous experiments).

The sequential testing procedure was planned in R using the *getDesignGroupSequential* function from the *rpact* package (the script used to plan the study is provided on the repository site). It was decided to realize one interim look after the collection of 70% of the data. The chosen option for the alpha spending function controlling for the Type-I error rate in a sequential testing procedure was *asP* (Pocock type alpha spending). The analysis revealed critical (one-sided) p-values for the planned interim look at $0.7N$ and the final look at $1.0N$ of $p_{interim} = .039$ and $p_{final} = .027$, respectively. Thus, an empirical p at the planned interim look smaller than $p_{interim}$ would mean that data collection can be terminated after this interim look. Using the *getSampleSizeMeans* function from the *rpact*, the final sample size needed to yield 80% test power was $N = 789$, which was rounded up to $N = 800$. This meant that the planned interim look at $0.7N$ had to take place after $N = 0.7 \cdot 800 = 560$ ($n = 70$ in each of the in total eight conditions). As the planned t-test was indeed significant at this point, data collection of the present study was stopped at this point.

8.1.2. Design, materials, and procedure

Whether subjects received information about low feature rates or not was manipulated between subjects (information about low feature rates: *presented* vs. *omitted*). Which of the two features served as the unobserved feature (unobserved feature: *beak* vs. *feet*) in the narrow latent scope bias test case was counterbalanced between subjects (see Fig. 11). A second counterbalancing factor manipulated between subjects was the orientation of the rating scale for that probe (side of narrow-scope explanation on the rating scale: *right* vs. *left*).

Subjects read a fictitious scenario about color abnormalities that biologists detected in the duck population of a Northern Italian lake. Some ducks were found to have blue beaks, blue feet, or both. These color abnormalities were the result of different independent gene mutations, called PIX67 (a mutation causing both a blue beak and blue feet), TOX20 (a mutation causing only a blue beak), and NAX20 (a mutation causing only blue feet). Together with the short scenario description, subjects also saw the illustration shown in Fig. 10. The main scenario description read:

> *Please read the following fictitious scenario thoroughly:*
>
> Biologists have noticed a peculiar phenomenon in the duck population of lake Caldazzo, a remote mountain lake in Northern Italy: Some ducks of lake Caldazzo happen to have blue feet (instead of the typical orange), others happen to have blue beaks (instead of the typical yellow), and yet others happen to have both blue feet and blue beaks.
>
> It was soon found out that these abnormalities are caused by genetic mutations existing in some ducks of the duck population of lake Caldazzo.
>
> Three mutations explain the observed color deviations:
>
> - A mutation of a gene called PIX67, leading to a blue beak and blue feet.
> - A mutation of a gene called TOX20, leading only to a blue beak.
> - And a mutation of a gene called NAX20, leading only to blue feet.
>
> *A graphic summary of what these mutations do is given below. Please study this information carefully. Your understanding of these mutations and their effects on duck color is crucial for the present study.*

After this initial information, subjects proceeded to a new screen where they received additional information about the mutations. Subjects read:

The biologists managed to collect DNA samples from all the ducks of Lake Caldazzo. A laboratory analysis of these samples revealed that all three mutations occur equally often in the population. That means each duck has the same chance of having the PIX67 mutation (causing both blue beak and blue feet), the TOX20 mutation (causing only a blue beak), or the NAX20 mutation (causing only blue feet). Also, the mutations occur independently of each other, which means that having one of the mutations has no influence on the probability of having another mutation. The analyses the biologists conducted also showed that 5% of the ducks in the population have blue feet [a blue beak].

The feature mentioned in the last sentence of the description about the feature rate was the one that served as the unobserved (latent) feature in the test phase. For subjects in the condition without prevalence information, the last sentence about the feature rate was omitted.

Subjects had to pass a comprehension test in which they answered different multiple-choice questions. The first tested subjects' knowledge of the different effects of the three different mutations. The second tested if they understood what it means that the mutations are independent. A third question tested if subjects understood that all three mutations are equally frequent in the population. Subjects in the condition with explicit feature rate information had to answer an additional question, which asked them to select the correct prevalence (5%) of the feature (blue beak vs. blue feet) that served as the unobserved (latent) feature in the test phase.

Subjects who failed the test were led back to the instructions and then got a new chance to pass the test. The data of subjects who needed more than three attempts were excluded from all analyses. Subjects who passed the comprehension test proceeded to the test phase. They were informed that they were going to see a number of pictures of ducks that were captured by a wildlife photo trap. They also learned that they were going to be asked for the most probable explanation of the ducks' appearances.

The first part of the test phase probed the narrow latent scope bias. Subjects were shown one of the two possible narrow latent scope bias test pictures (cf. Fig. 11). A prompt above the picture read: "The camera trap that was installed at the lakeside took the following picture of a duck that passed by". Subjects were asked the following test question: "Based on what you've learned, what is the more probable explanation for this duck's appearance?" Responses were given on an eleven-point rating scale with the endpoints labeled *Definitely a TOX20 [NAX20] mutation (blue beak only) [blue feet only]* and *Definitely a PIX67 mutation (blue beak and feet)*, and the midpoint labeled *Both equally likely*.[7] Whether the narrow or the broad-scope explanation was on the left or the right endpoint of the scale (scale orientation) was counterbalanced between subjects. On a subsequent screen, subjects were also asked to write brief explanations.

The second part of the test phase tested subjects' explanatory simplicity preference. The relevant test picture showed a duck that had both a blue beak and blue feet. Subjects answered the following question: "Based on what you've learned, what is the most probable explanation for this duck's appearance?" Subjects selected an answer from a list of the following seven possibilities (this test question format was adopted from experiments of Lombrozo, 2007):

- This duck has PIX67 mutation (blue beak and blue feet) and a TOX20 mutation (blue beak).
- This duck has a PIX67 mutation (blue beak and blue feet).
- This duck has a TOX20 mutation (blue beak).
- This duck has a PIX67 mutation (blue beak and blue feet) and NAX20 mutation (blue feet).
- This duck has a NAX20 mutation (blue feet).
- This duck has no mutation affecting beak and feet color.
- This duck has a TOX20 mutation (blue beak) and a NAX20 mutation (blue feet).

A preference for explanatory simplicity would be expressed by subjects who select the PIX67 mutation, as this is a single (common cause) mutation accounting for both color abnormalities (cf. Fig. 1).

This second phase also contained three additional control pictures that probed subjects' understanding of the scenario. Subjects were asked the same test question and had to choose an option from the same list. These three additional control pictures were: a picture of a duck with unaltered beak and feet, a picture of a duck with only a blue beak, and one showing a duck that had only blue feet. All four pictures in this second part of the test phase were presented in random order. The order of the response options was randomized between subjects but every participant saw the same order for all four test pictures of that part of the test phase.

8.2. Results and discussion

8.2.1. Subjects' ratings for the narrow latent cope bias probe

The results for the narrow latent scope bias in the different feature rate information conditions are shown in Fig. 13. Fig. 13a shows subjects' explanation ratings and Fig. 13b shows a classification of these ratings. Replicating previous experiments in this paper that used a rating scale, overall only small narrow latent scope biases were found. This was true in both feature rate information conditions. In the condition in which subjects did not receive information abut the base rate of the unobserved feature, the mean of subjects' ratings was close to the normative midpoint of the scale, $M = -0.04$, and its 95% CI included that midpoint, 95% CI $[-0.19, 0.10]$. In the condition in which subjects were given information about a low base rate of the unobserved feature, the mean indicated a small narrow latent scope bias. The mean was $M = -0.275$, and its CI excluded the normative midpoint of the scale, 95% CI $[-0.45, -0.11]$. A directed Welch two-sample t-test comparing the means in the two condition was significant, $t(544.1) = 2.09$, $p = .019$, $d = 0.18$.

Fig. 13b shows that only a minority of participants in both conditions preferred the narrow latent scope explanation, however. In the condition without feature rate information, 5.71% (95% CI $[0.03, 0.09]$) preferred the narrow latent scope explanation. This proportion was higher in the condition in which subjects had learned that the latent feature is rare. Here, 12.10% (95% CI $[0.08, 0.16]$) indicated that the narrow latent scope explanation was more likely. A directed 2-sample test for equality of proportions showed that this difference was significant, $\chi^2(1) = 7.12$, $p = .004$ (one-sided).

This experiment aimed to minimize the potential influence of pragmatic reasoning but at the same (in one of the conditions) time made the low overall probability of the unobserved feature particularly salient to participants, the factor that should increase narrow latent scope biases according to the inferred evidence account. Most subjects were not led astray even under this condition and still responded normatively correct. However, the subgroup of subjects who preferred the narrow latent scope explanation, for the lack of pragmatic reasons, may indeed have been influenced by the low overall probability of the unobserved feature. These subjects thus seem to have committed a genuine reasoning error.

8.2.2. Subjects' explanations

Table 5 summarizes relevant explanation categories that were identified. The reported test results come from 2-sample equality of proportion tests comparing the two feature rate information conditions. As in

[7] I'd like to thank an anonymous reviewer for pointing out that a clarification might be in order here. The structure of the scenario deviates from those of previous scenarios. Unlike in previous scenarios, the different possible explanations in the present scenario are not mutually exclusive; both could be true at the same time, or even neither of them might be correct. Yet, a question asking which of two possible explanations is more probable is still a valid question.

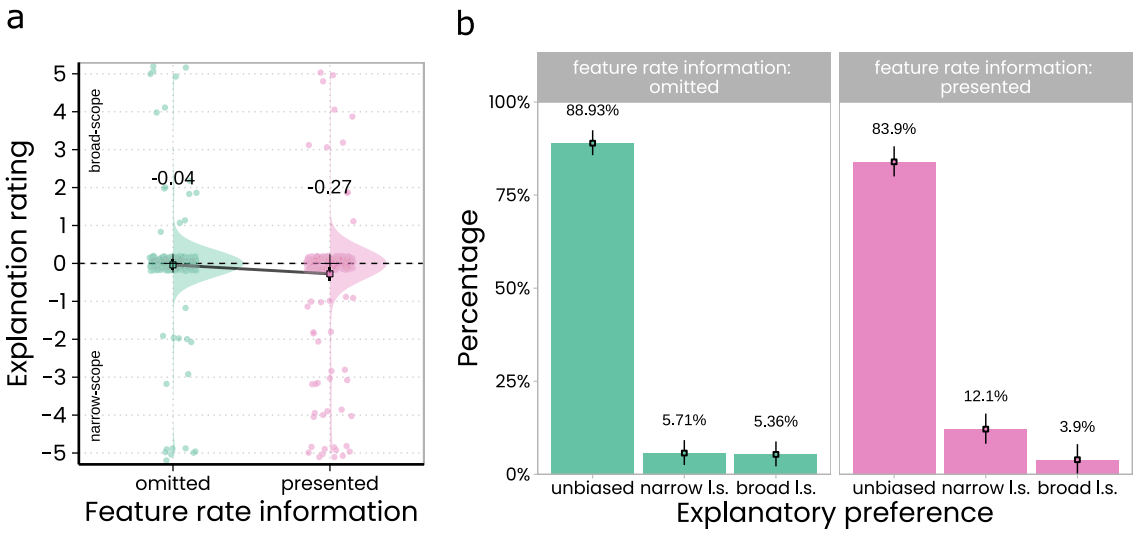

Fig. 13. Subjects' ratings for the unobserved feature probe in Experiment 3.
Note. a: Squares and annotations denote means, "+" denote medians. Error bars represent 95% CIs of the means. Jittered dots show subjects' individual ratings, and the density plot their distribution. b: Figure panels represent the two feature rate information conditions. Proportions of subjects having certain explanatory preferences as indicated by their ratings. Error bars represent 95% CIs of the proportions, which were computed in R using the "MultinomCI" function from the "DescTools" package with the default estimation option "sisonglaz".

Table 5
Relevant subject explanation categories identified in Exp. 3.

Explanation category	n (%) in "feature rate information: omitted"	n (%) in "feature rate information: presented"	Result of proportion test
Clearly stating that both explanations are equally likely	171 (61%)	177 (63%)	$p = .60$ (two-sided)
Clearly stating that low feature base rate was the relevant aspect	0 (0%)	12 (4%)	$p < .001$ (one-sided)

previous experiments, the explanations of subjects who gave unbiased ratings tended to point out the statistical facts of the scenario. More than two thirds of the explanations clearly described that both possible mutations are equally likely to explain the test case's appearance. Examples are: "It is evident in the picture that the duck had a blue beak, hence it could either be a PIX67 or TOX20, and since both mutations have an equal chance of occurrence it follows that the duck in the picture could have any of the mutations" and "Since the mutations vary independently of each other and are distributed equally among the population. There is an equal probability that the mutation is PIX67 and NOX20". The proportions of correct explanations did not differ between the feature rate information conditions, $\chi^2(1) = 0.27$, $p = .60$ (two-sided).

Explanations that clearly described the process leading to genuine narrow latent scope biases according to the inferred evidence account were provided by 4% of the subjects. These subjects were all in the condition in which subjects received explicit information about a low rate of the unobserved feature. This proportion was significantly higher than in the condition in which no feature rate information was provided, $\chi^2(1) = 12.26$, $p < .001$ (one-sided). Example explanations are: "Since birds with blue feet are only 5% of the population I took that into account when deciding the rating probability I did", "Only 5% of the population have blue feet so it is highly likely that this duck has only got a blue beak and the mutation TOX20", and "There is only a 5% chance of the duck having a blue beak so its more than likely this duck does not have a blue beak". Crucially, subjects who wrote this kind of explanation also indicated a preference for the narrow latent scope explanation. This finding shows that the reasoning process assumed by the inferred evidence account can, at least sometimes, lead to unwarranted narrow latent scope preferences.

In the condition in which subjects did not receive explicit feature base rate information, another question was what kind of explanations those (6% of the) subjects wrote who preferred the narrow latent scope mutation. These subjects mostly wrote explanations that did not allow a clear answer as to whether they came to their conclusion based on a fallacious reasoning process or not. Examples are: "because i only saw the beak of the duck i chose TOX but again it could be pix67 if the feet are blue", "the duck has blue feet but im not sure of the beak so i guessed". One subject wrote an explanation that quite clearly documented a fallacious conclusion, although it was not clear whether that process was the one assumed by the inferred evidence account: "I know they all occur equally, but we know for a fact that it has blue feet already, so I thought that tipped he scale a bit".

Finally, two explanations could be found that reported on an initial impulse to select the narrow-scope explanation: "At first I would of gone for 100% NAX20, yet the ducks head is under water and may have an orange or blue beak so both are equally likely. Also if one duck has one mutation, it's equally likely that it could have another mutation aswell" and "You can clearly see the ducks beak, which would instantly put this duck as a TOX20 mutation, however, the feet are under the water so you cannot see if they are blue or not, so you can't tell if it's a TOX20 or a PIX67". These explanations are interesting because they suggest that subjects' tendency to commit narrow latent scope biases might increase if they were forced to respond fast.

In sum, the results revealed at best a small tendency to commit narrow latent scope biases, even if information about a low overall probability of the unobserved feature is explicitly provided. Yet, at least some subjects seem to have relied on this information, which then led them to commit a genuine narrow latent scope bias. The results also suggest that in contexts in which feature base rates are less salient, or not mentioned at all (like in the no feature rate information condition of the present study or scenarios like the Tokolo tribe vignette), reasoners' tendency to be influenced by them seems rather unlikely.

8.2.3. *Subjects' ratings for the simplicity probe*

Another goal of this study was to see if reasoners' preference for simpler over more complex explanations (see, e.g., Lombrozo, 2007)

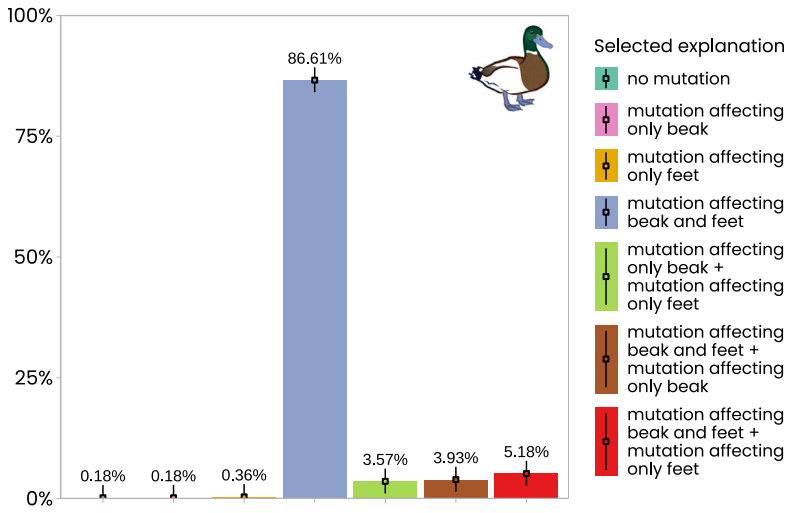

Fig. 14. Subjects' explanatory preferences for the simplicity preference probe in Experiment 3.
Note. Error bars represent 95% CIs of the proportions, which were computed in R using the "MultinomCI" function from the "DescTools" package with the default estimation option "sisonglaz". (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

could be replicated, which would provide converging evidence that reasoners explanatory preferences tend to be normative.

The results for the simplicity preference probe (a test picture of a duck with both blue feet and a blue beak) are shown in Fig. 14. Based on previous studies documenting a preference for explanatory simplicity, the prediction was that subjects would tend to say that the mutation that causes both observed effects is the most probable explanation. As can be seen in Fig. 14, this was the case. The mutation causing both mutations was selected as the most probable explanation by 87% (95% CI $[0.84, 0.89]$) of participants, while only 13% selected a different explanation.

In addition to the simplicity probe, subjects also rated a picture of a duck with "normal" feet and beak, a picture of a duck with only a blue beak, and a picture of a duck with only blue feet. Subjects' responses to these probes can be regarded as attention checks. For the test picture showing a duck with unaltered beak and feet, 98% (95% CI [0.97, 0.99]) of the subjects said that the most probable explanation was *no mutation*. For the test picture showing a duck with a blue beak, 96% (95% CI $[0.95, .98]$) of the subjects said that the most likely explanation was the mutation that causes only this effect. Similarly, 98% (95% CI $[0.97, 0.99]$) of the subjects said that the mutation leading only to blue feet was the most likely explanation for the test picture showing a duck with blue feet. The results for these three additional test pictures show that subjects correctly understood the scenario.

8.2.4. Conclusion

The results of this study complement the findings of this paper's previous studies: All in all, subjects tended to give normatively correct answers. Most said that both explanations are equally likely to be true. Only a minority tended to favor a narrow latent scope explanation. Some of these subjects seem to have been influenced by information about a low rate of the unobserved feature, which shows that reasoners, at least sometimes, engage in the erroneous reasoning process assumed by the inferred evidence account. Yet, the effect of feature rate information was weak. One reason might be that the information was not salient enough. This seems unlikely, though. The information was presented last in the scenario description and subjects' knowledge about it was probed in a separate comprehension test question. It seems more plausible that most subjects understood that this information was irrelevant.

In the second part of this experiment that tested explanatory simplicity preferences, it was found that most subjects indeed preferred the single factor that accounted for both observed features. This result replicates earlier findings of a preference for explanatory simplicity. In sum, the results of this study document that reasoners tend to explain reasonably.

9. General discussion

Previous studies on lay people's explanatory reasoning have yielded mixed results. On the one hand, several studies suggest that reasoners have explanatory preferences that reflect the same explanatory virtues endorsed by philosophers and scientists. The picture that these studies paint of human explanatory reasoning is one where, by and large, reasoners follow normative principles (Lombrozo & Vasilyeva, 2017, see also), even though they might not always be completely accurate.

A different picture has been painted by a series of studies (Johnson et al., 2016; Johnston, Johnson, Koven, & Keil, 2017; Khemlani et al., 2011; Sussman et al., 2014) that investigated reasoners' explanatory preferences in situations where pieces of (diagnostic/relevant) evidence are missing (latent). A seemingly robust deviation from normativity has been observed in these cases, called the "narrow latent scope bias": if the status of relevant pieces of evidence is unspecified, reasoners seem to prefer explanations that do not predict these latent pieces of evidence – even if they have learned that the potential explanations have identical prior probabilities and predict the manifest evidence equally well, i.e., are objectively equally probable in the target situation.

The present paper revisited the narrow latent scope bias. Its results suggest that the picture of a pronounced, robust, narrow latent scope bias in explanatory reasoning requires some correction. The main conclusion that can be drawn is that the bias is less robust than previously thought. A first finding is that a strong narrow latent scope bias only occurs when subjects are forced to commit to a wrong answer, but not when they are allowed to respond correctly. As soon as they are allowed to do so, the vast majority of subjects in the present experiments did; and only a minority of participants continued to favor narrow latent scope explanations. This result is in line with what has also been observed in another recent experiment by Tsukamura et al. (2022). The authors used a medical test scenario similar to those found in Johnson et al. (2016) and allowed their subjects to respond on a continuous slider. Like in the present experiments, Tsukamura et al. (2022) found that narrow latent scope biases occurred only in a subgroup of participants.

A central finding of this paper is that subjects' preferences for narrow latent scope explanations are nuancedly influenced by pragmatic reasoning. The paper looked at two pragmatic factors, feature

diagnosability and ambiguity of the test query ("more satisfying" vs. "more probable"). In experimental scenarios where features differ with respect to diagnosability and in scenarios with ambiguous test queries, it was found that subjects who prefer narrow latent scope explanations tend to have rational reasons for doing so.

Although the influence of test query formulation (probability vs. satisfaction) was relatively small in the present studies, it still seems warranted to recommend that future studies use formulations that are as unambiguous as possible. As the narrow latent scope bias is considered a non-normative bias for probabilistic reasons (it is considered a bias because it deviates from the competing explanations' posterior odds ratio), test questions probing the bias should use probabilistic terminology. A more general conclusion that can be drawn from the present findings is that a potential influence of pragmatic factors ought to be considered whenever researchers use verbal scenarios and test questions to probe reasoning. As has been noted by Woodward (2021): "The experimenter uses certain words in the probe but cannot control how those words are interpreted by the subjects, so that subjects may be answering a different question or engaged in a different task than what the experimenter intended" (p. 339).

Evidence for a nuanced influence of pragmatic reasoning was obtained not only from subjects' responses to test queries asking them to make a forced choice or to provide a rating on a scale, but also from brief explanations that subjects wrote. While asking subjects to explain why they did what they did might not be helpful in every kind of psychological experiment, in the context of the present paper subjects' explanations have proven insightful. Thus, at least in studies investigating how people reason about problems whose relevant aspects are verbally conveyed in the description of a test scenario, asking subjects to also explain their behavior may be considered a worthwhile option.

This paper's finding of a nuanced influence of pragmatic reasoning in people's evaluation of competing explanations is interesting in another respect. There has been a debate over how sophisticated, versus how reliant on heuristics, people are in their explanatory reasoning (see, e.g., Dellsén, 2018; McGrew, 2003). For example, differences in explanatory simplicity often seem to play the role of a heuristic indicator signaling which explanation is more likely. In line with this view, reasoners have been found to overapply this rule of thumb (see, e.g., Lombrozo & Vasilyeva, 2017): they sometimes seem to need disproportionate (probabilistic) evidence before letting go of simpler and opting for more complex explanations (see, e.g., Experiments 2 and 3 in Lombrozo & Vasilyeva, 2017). Similarly, previous findings of robust narrow latent scope biases could be interpreted as evidence for the overapplication of the heuristic to prefer explanations that better account for evidence to cases where that evidence is only inferred rather than known (see, e.g., Johnston et al., 2017). The present studies, by contrast, suggests that this largely does not occur. In fact, it has been found that narrow latent scope preferences often seem to track sophisticated probabilistic inferences (e.g., based on pragmatic factors such as feature diagnosability). This reduces the evidence for the heuristic view of explanatory reasoning, and opens up the question of whether other apparent biases/evidence for overapplied heuristics might also be driven by sophisticated probabilistic inferences based on pragmatic/contextual factors. Indeed some past work has already provided some evidence for this. For example, based on what they found in supplementary study (Vrantsidis & Lombrozo, 2022) suggest that some of the applications of simplicity are quite nuanced, and perhaps justifiable for pragmatic reasons. One of their findings was that scenario wording influenced participants' assumptions about the conditional independence of the effects predicted by competing explanations, and thus the direction of explanatory simplicity effects. Similarly, work by Zemla and colleagues (see, e.g., Sloman, Zemla, Lagnado, Bechlivanidis, & Hemmatian, 2019; Zemla et al., 2017, 2023) may be taken to suggest that the degree to which reasoners adhere to different explanatory virtues (e.g., simplicity or abstractness) also may depend on the pragmatic "purpose" (Sloman et al., 2019, p. 14) an explanation is assumed to have; they contrast the example of a policymaker who might favor more abstract and generalizable explanations with one of a private investigator who might demand as much detail about a case as possible). Future work may continue looking into this.

Although this paper shows a clear influence of pragmatic reasoning on the narrow latent scope bias, it is not the first to consider the possibility of an influence of pragmatic reasoning on the narrow latent scope bias. For example, Johnson et al. (2016) report experiments in which they sought to examine the influence of pragmatic reasoning. In one experiment, the authors aimed to "block pragmatic interpretations of the speakers' claim to ignorance [about the unobserved feature in the test scenario]" (p. 50). In this study, the scenario was a fictitious medical scenario about two diseases causing abnormal levels of different blood substances. The test case described a patient who definitely had one of the two diseases. Blood tests confirmed the presence of the shared (undiagnostic) symptom. To block pragmatic inferences, the authors included a reason for why the status of the diagnostically relevant symptom remained unknown. Subjects learned that the status of the unique symptom was still unclear because the results had not yet come back from the laboratory. The experiment yielded the predicted narrow latent scope bias. In a follow-up experiment, the authors aimed to "measure" (p. 53) the impact of pragmatic reasoning by directly contrasting this condition (the explanation condition) with one in which they used their original description of the test case (no explanation condition). In this no explanation condition, the test case description simply read "You don't know whether the patient's [blood substance] levels are normal or abnormal". The results revealed narrow latent scope biases in both conditions and their magnitude did not differ, which led the authors to conclude that "pragmatic inferences [...] seem to have modest influences at most, for the stimuli used in these experiments". (p. 55).

One possibility for why no obvious influence of pragmatic reasoning could be revealed between the explanation and no explanation conditions in Johnson et al.'s (2016) studies is that the explanation condition actually did not block pragmatic inferences. For example, one possibility is that subjects might have thought that there must be a reason why the tests for the first blood substance came back from the lab but not those for the second. If a plausible reason can be found for why the unobserved feature is absent, it seems warranted to say that the narrow-latent scope explanation is more probable. For example, in the given medical scenario we might imagine that initial rapid tests yielded a positive result only for the manifest substance, which is why additional tests must be conducted on the unobserved substance. One might plausibly conclude that the unobserved substance is likely absent in this case.

To successfully block pragmatic inferences the reason for why the latent symptom's status is unknown must be perceived by subjects as being unrelated to the symptom's actual status. The repository site of the present paper reports on a supplementary study in which this was tested (see https://simonstephan31.github.io/revisit_nlsbias/expSup_mat.html). In this supplementary experiment (following Johnson et al., 2016, Exp. 3), two conditions that were contrasted (of three conditions in total) were a no explanation condition and a novel explanation condition. In the explanation condition, the reason for why the unobserved feature's status remained unknown that was presented could not plausibly be linked to the feature's actual status. Subjects read: "[...] you ordered blood tests for the patient. The letter with the results has just come back. Unfortunately, you accidentally spill your cup of coffee over the document and now only part of it is still readable. From what is still readable, you can see that Patient #890 indeed has abnormal levels of Lian. However, due to a big stain of coffee right where the results for the Gludon levels are printed, you cannot see whether or not the Patient #890 also has abnormal levels of Gludon.". The mean narrow latent scope bias found in this condition was close to and not significantly different from zero ($M = -0.167$,

95% CI [−0.48, 0.10]). Also, the proportion of subjects showing a narrow latent scope bias in the novel explanation condition was less than half (12%) of what it was in the no explanation condition (29%). This supplementary study adds to the findings of the main experiments of this paper: it is another demonstration that pragmatic reasoning plays a role in subjects' preferences for narrow latent scope explanations. It also suggests that pragmatic inferences can be blocked. If this happens, almost no bias is observed.

Although the narrow latent scope biases measured in the present experiments was at best weak (if subjects were allowed to respond correctly) and influenced by pragmatic factors, the studies of the present paper also provided some evidence for the reasoning process postulated by the inferred evidence account (Johnson et al., 2016). However, the influence of that reasoning process seems to be relatively weak and restricted to situations where low base rates of the effects/features are made explicit. Among those subjects in Experiments 2a, b, and 3 who preferred the narrow latent scope explanation, some of their provided explanations indicated that they ruled the broad latent scope explanation out because they believed that the unobserved feature, due to its low prior probability, would probably be absent in the test case. This minority of subjects thus exhibited an explanatory preference that actually resulted from a genuine statistical reasoning error.

A question that arises in light of the present results is what significance the inferred evidence account still has? The general idea of the account is that "people perform explanatory reasoning using not only the observed evidence, but also inferred evidence (Johnson, Rajeev-Kumar, & Keil, 2014). That is, when some evidence is unavailable but potentially diagnostic, people make a guess as to what that evidence would be, if it were known" (p. 43). The present paper provides clear evidence for this general idea of the inferred evidence account, but not for the specific fallacious reasoning process assumed to implement it: In the experiments that tested the role of feature diagnosability, subjects who indicated a preference for narrow latent scope explanations tried to infer the status of the unobserved feature and concluded that it is probably absent. This finding is in line with the general idea of the inferred evidence account. Importantly, though, subjects tended to provide rational/reasonable justifications. For example, in the original Tokolo scenario, subjects tended to say they thought the fishing net was probably absent because it should be (easily) visible if one is close enough to see the spear. This finding goes against the inferred evidence account, which postulates that what lies behind the general notion of "when some evidence is unavailable but potentially diagnostic, people make a guess as to what that evidence would be, if it were known" be a fallacious probabilistic reasoning process: it is assumed that reasoners would erroneously infer the absence of the unobserved feature because of its low base rate. The view of a strong impact of this specific fallacious reasoning process is not well supported by the present studies, as it was found to produce genuine narrow latent scope biases at best sometimes.

Another question is whether pragmatic reasoning and the reasoning process postulated by the inferred evidence account fully explain the narrow latent scope bias. The results of the present paper suggest that they explain at least a lot of it. In Experiment 3, where pragmatic reasoning was blocked and no explicit information about feature base rates was provided, the narrow latent scope preferences nearly completely disappeared. In the condition in which subjects were given information about low feature rates, a small bias occurred and subjects explanations revealed that they tended to show it because of that information. That said, the present paper revisited the narrow latent scope bias only in adult human reasoners, but it has also been probed and found in human children (see, e.g., Johnston et al., 2017). While present studies do not support the view of a strong, robust, narrow latent scope bias in adults, and at the same time document that much of it can be explained by pragmatic reasoning, it is still possible that a stronger bias exists in children. In fact, the experiments testing children seem better controlled with respect to pragmatic factors than many experiments that tested adult subjects. In part, this is because these studies, like Experiment 3 in the present paper, relied more on visual than on verbal presentations of the test situation (but see also Experiment 2 in Sussman et al., 2014). An example are Johnston et al.'s (2017) Experiments 3 and 4, which used a fictitious machine producing two different effects (the turning on of a light and of a fan). Clear narrow latent scope preferences were observed there. Interestingly, though, the narrow latent scope preferences found in their Experiment 3 are unlikely to have resulted from the fallacious probabilistic reasoning process assumed by the inferred evidence account, because subjects had learned that the target effects occur with a 50% probability. It is possible that subjects still inferred the absence of the latent effect in these studies (which is what the authors also think). Future studies might further investigate why. All in all, findings on the narrow latent scope bias in children, together with the present ones, leave open the possibility that a stronger tendency to commit the narrow latent scope bias might exist in childhood, which then decreases during development. Such a developmental trajectory has already been documented for other explanatory reasoning biases (see, e.g., Cimpian & Steinberg, 2014). It would also fit with some of subjects' explanations observed in the present Experiment 3, in which subjects described that they had to refrain from giving in to an initial impulse to select the narrow-scope explanation. Such an initial impulse might also be contributing to the stronger narrow latent scope preferences observed under a forced choice response format.

Finally, the present paper is also the first to provide a detailed insight into the distribution of the narrow latent scope bias. Knowing the distribution of a behavior may generally be regarded relevant, but it seems to be particularly important in cases where a documented behavior is claimed to systematically deviate from a normative standard. In cases like the narrow latent scope bias, where a behavior is non-normative on the group level, it is relevant to know whether it occurs in the majority of reasoners or whether it is driven by only a subgroup. The studies of this paper provide a clear answer: in all conditions of the present paper's experiments that allowed subjects to give the correct answer, the vast majority of subjects responded normatively correct. Group-level deviations from the normatively correct answer in all these experiments resulted from a small number of the participants. This was true even in conditions where a preference for narrow-scope explanations could be justified for pragmatic reasons, either because the unobserved feature should have been as easy to identify as the manifest feature if it had really been present in the target situation (feature diagnosability), or because the test question was ambiguous (satisfaction vs. probability). Also, this was true even in those experiments (Experiments 2a, 2b, and 3) in which subjects were given explicit information about a low feature base rate, which should increase the narrow latent scope bias according to the inferred evidence account (Johnson et al., 2016). Even in Experiment 3, which made low feature base rates particularly salient, only a subgroup of subjects deviated from the normatively correct answer and preferred the narrow latent scope explanation.

10. Conclusion

The studies in this paper show that, by and large, reasoners have normative explanatory preferences in both latent scope as well non-latent scope situations. Human reasoning is not always accurate and certain robust reasoning biases undoubtedly exist. The narrow latent scope bias might not be one of them, however.

CRediT authorship contribution statement

Simon Stephan: Conceptualization, Methodology, Software, Data curation, Writing – original draft, Visualization, Investigation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

All data and analyses scripts have been made publicly available on a GitHub repository, which can be accessed via a GitHub page at https://simonstephan31.github.io/revisit_nlsbias/index.html.

Acknowledgments

I would like to thank Julia Larysch for her help in developing some of the experiments. I would also like to thank Sarah Placì for helpful comments, and two anonymous reviewers for the very constructive evaluation of this work. During the final stage of this project the author was funded by a Reinhart Koselleck project (WA 621/25-1), funded by the Deutsche Forschungsgemeinschaft (DFG).

References

Bonawitz, E. B., & Lombrozo, T. (2012). Occam's rattle: children's use of simplicity and probability to constrain inference. *Developmental Psychology*, *48*(4), 1156–1164.

Brewer, W. F., Chinn, C. A., & Samarapungavan, A. (1998). Explanation in scientists and children. *Minds and Machines*, *8*, 119–136.

Cheng, P. W., & Lu, H. (2017). Causal invariance as an essential constraint for creating a causal representation of the world: Generalizing the invariance of causal power. In M. R. Waldmann (Ed.), *The Oxford handbook of causal reasoning* (pp. 65–84). New York: Oxford University Press.

Cimpian, A., & Steinberg, O. D. (2014). The inherence heuristic across development: Systematic differences between children's and adults' explanations for everyday facts. *Cognitive Psychology*, *75*, 130–154.

Corriveau, K. H., & Kurkul, K. E. (2014). "Why does rain fall?": Children prefer to learn from an informant who uses noncircular explanations. *Child Development*, *85*, 1827–1835.

Cumming, G. (2013). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge.

de Leeuw, J. R., Gilbert, R. A., & Luchterhandt, B. (2023). jsPsych: Enabling an open-source collaborative ecosystem of behavioral experiments. *Journal of Open Source Software*, *8*, 5351.

Dellsén, F. (2018). The heuristic conception of inference to the best explanation. *Philosophical Studies*, *175*, 1745–1766.

Fernbach, P. M., Darlow, A., & Sloman, S. A. (2011). Asymmetries in predictive and diagnostic reasoning. *Journal of Experimental Psychology: General*, *140*, 168–185.

Fernbach, P. M., & Rehder, B. (2013). Cognitive shortcuts in causal inference. *Argument & Computation*, *4*, 64–88.

Gopnik, A. (2000). Explanation as orgasm and the drive for causal knowledge: The function, evolution, and phenomenology of the theory-formation system. In F. Keil, & R. A. Wilson (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 299–324). Cambridge: MIT Press.

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

Hahn, U., & Harris, A. J. (2014). What does it mean to be biased: Motivated reasoning and rationality. *Psychology of Learning and Motivation*, *61*, 41–102.

Hitchcock, C. (2009). Causal modelling. In H. Beebee, C. Hitchcock, & P. Menzies (Eds.), *The Oxford handbook of causation* (pp. 299–314). New York: Oxford University Press.

Johnson, S. G., Johnston, A., Toig, A., & Keil, F. (2014). Explanatory scope informs causal strength inferences. In P. Bello, M. Guarini, & B. Scassellati (Eds.), *Proceedings of the 36th annual conference of the cognitive science society*. 2453–2458.

Johnson, S. G., Rajeev-Kumar, G., & Keil, F. (2014). Inferred evidence in latent scope explanations. In P. Bello, M. Guarini, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 707–712).

Johnson, S. G., Rajeev-Kumar, G., & Keil, F. C. (2016). Sense-making under ignorance. *Cognitive Psychology*, *89*, 39–70.

Johnston, A. M., Johnson, S. G., Koven, M. L., & Keil, F. C. (2017). Little Bayesians or little Einsteins? Probability and explanatory virtue in children's inferences. *Developmental Science*, *20*, Article e12483.

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

Khemlani, S. S., Sussman, A. B., & Oppenheimer, D. M. (2011). Harry Potter and the sorcerer's scope: latent scope biases in explanatory reasoning. *Memory & Cognition*, *39*, 527–535.

Lagnado, D. (1994). *The psychology of explanation: A Bayesian approach* (Unpublished doctoral dissertation), Masters Thesis. Schools of Psychology and Computer Science, University of ....

Lakens, D. (2022). Improving your statistical inferences. Retrieved from https://lakens.github.io/statistical_inferences/. https://doi.org/10.5281/zenodo.6409077.

Lim, J. B., & Oppenheimer, D. M. (2020). Explanatory preferences for complexity matching. *PLoS One*, *15*(4), Article e0230929.

Lipton, P. (2004). *Inference to the best explanation* (2nd ed.). New York, NY: 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*(3), 232–257.

Lombrozo, T. (2009). Explanation and categorization: How "why?" informs "what?". *Cognition*, *110*, 248–253.

Lombrozo, T. (2010). Causal–explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, *61*, 303–332.

Lombrozo, T. (2011). The instrumental value of explanations. *Philosophy Compass*, *6*(8), 539–551.

Lombrozo, T. (2012). Explanation and abductive inference. In K. Holyoak, & R. G. Morisson (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 260–276). New York: Oxford University Press.

Lombrozo, T. (2016). Explanatory preferences shape learning and inference. *Trends in Cognitive Sciences*, *20*(10), 748–759.

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

McGrew, T. (2003). Confirmation, heuristics, and explanatory reasoning. *British Journal for the Philosophy of Science*, *54*, 553–567.

Meder, B., & Mayrhofer, R. (2017). Diagnostic causal reasoning with verbal information. *Cognitive Psychology*, *96*, 54–84.

Meder, B., Mayrhofer, R., & Waldmann, M. R. (2014). Structure induction in diagnostic causal reasoning. *Psychological Review*, *121*(3), 277–301.

Pacer, M., & Lombrozo, T. (2017). Ockham's razor cuts to the root: Simplicity in causal explanation. *Journal of Experimental Psychology: General*, *146*, 1761–1780.

Paul, L. A., & Hall, E. J. (2013). *Causation: A user's guide*. New York: Oxford University Press.

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

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

Preston, J., & Epley, N. (2005). Explanations versus applications: The explanatory power of valuable beliefs. *Psychological Science*, *16*, 826–832.

Read, S. J., & Marcus-Newhall, A. (1993). Explanatory coherence in social explanations: A parallel distributed processing account. *Journal of Personality and Social Psychology*, *65*, 429–447.

Shimojo, A., Miwa, K., & Terai, H. (2020). How does explanatory virtue determine probability estimation?—Empirical discussion on effect of instruction. *Frontiers in Psychology*, *11*, Article 575746.

Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. S. (2022). afex: Analysis of factorial experiments. [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=afex (R package version 1.1-1).

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

Sloman, S., Zemla, J., Lagnado, D., Bechlivanidis, C., & Hemmatian, B. (2019). Are humans intuitive philosophers? In S. R. Grimm (Ed.), *Varieties of understanding* (pp. 231–250). Oxford: Oxford University Press.

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, prediction, and search*. New York, NY: Springer-Verlag.

Sussman, A. B., Khemlani, S. S., & Oppenheimer, D. M. (2014). Latent scope bias in categorization. *Journal of Experimental Social Psychology*, *52*, 1–8.

Thagard, P. (1993). *Conceptual revolutions*. Princeton: Princeton University Press.

Tsukamura, Y., Wakai, T., Shimojo, A., & Ueda, K. (2022). How does the latent scope bias occur?: Cognitive modeling for the probabilistic reasoning process of causal explanations under uncertainty. In J. Culbertson, A. Perfors, H. Rabagliati, & V. Ramenzoni (Eds.), *Proceedings of the 44th annual meeting of the cognitive science society* (pp. 818–823).

Vrantsidis, T. H., & Lombrozo, T. (2022). Simplicity as a cue to probability: multiple roles for simplicity in evaluating explanations. *Cognitive Science*, *46*(7), Article e13169.

Waldmann, M. R. (Ed.), (2017). *The Oxford handbook of causal reasoning*. New York: Oxford University Press.

Waldmann, M. R., Meder, B., von Sydow, M., & Hagmayer, Y. (2010). The tight coupling between category and causal learning. *Cognitive Processing*, *11*, 143–158.

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., & Rehder, B. (2011). Explaining drives the discovery of real and illusory patterns. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd annual meeting of the cognitive science society* (pp. 1352–1357).

Williams, J. J., Lombrozo, T., & Rehder, B. (2013a). Explaining drives the discovery of real and illusory patterns. In A. Nicholson, & P. Smyth (Eds.), *Proceedings of the 29th conference on uncertainty in artificial intelligence* (pp. 498–507).

Williams, J. J., Lombrozo, T., & Rehder, B. (2013b). The hazards of explanation: Overgeneralization in the face of exceptions. *Journal of Experimental Psychology: General*, *142*, 1006.

Wojtowicz, Z., & DeDeo, S. (2020). From probability to consilience: How explanatory values implement Bayesian reasoning. *Trends in Cognitive Sciences*, *24*, 981–993.

Woodward, J. (2021). *Causation with a human face: Normative theory and descriptive psychology*. Oxford University Press.

Zemla, J. C., Sloman, S., Bechlivanidis, C., & Lagnado, D. A. (2017). Evaluating everyday explanations. *Psychonomic Bulletin & Review*, *24*, 1488–1500.

Zemla, J. C., Sloman, S. A., Bechlivanidis, C., & Lagnado, D. A. (2023). Not so simple! Causal mechanisms increase preference for complex explanations. *Cognition*, *239*, Article 105551.