COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 42 (2018) 1345–1359

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12573

Bayesian Occam's Razor Is a Razor of the People

Thomas Blanchard,[a] Tania Lombrozo,[b] Shaun Nichols[c]

[a]*Department of Philosophy, Illinois Wesleyan University*
[b]*Department of Psychology, University of California – Berkeley*
[c]*Department of Philosophy, University of Arizona*

Received 24 October 2016; received in revised form 10 September 2017; accepted 25 September 2017

Abstract

Occam's razor—the idea that all else being equal, we should pick the simpler hypothesis—plays a prominent role in ordinary and scientific inference. But why are simpler hypotheses better? One attractive hypothesis known as Bayesian Occam's razor (BOR) is that more complex hypotheses tend to be more flexible—they can accommodate a wider range of possible data—and that flexibility is automatically penalized by Bayesian inference. In two experiments, we provide evidence that people's intuitive probabilistic and explanatory judgments follow the prescriptions of BOR. In particular, people's judgments are consistent with the two most distinctive characteristics of BOR: They penalize hypotheses as a function not only of their numbers of free parameters but also as a function of the size of the parameter space, and they penalize those hypotheses even when their parameters can be "tuned" to fit the data better than comparatively simpler hypotheses.

Keywords: Simplicity; Flexibility; Bayesianism; Probability; Explanation

1. Introduction

Occam's razor—the idea that all else being equal, we should pick the simpler hypothesis—plays a prominent role in ordinary and scientific inference (Baker, 2013; Lombrozo, 2016; Sober, 2015). But why are simpler hypotheses better? One attractive answer is the Bayesian Occam's razor (BOR), according to which Bayesian inference automatically penalizes hypotheses that are more complex in the sense that they contain more free parameters and/or free parameters with more possible values (Henderson, Goodman, Tenenbaum, & Woodward, 2010; Jefferys & Berger, 1992; MacKay, 2003; Rosenkrantz, 1977; see Sober, 2015, for critical discussion). Here, we examine whether people's

Correspondence should be sent to Thomas Blanchard, Department of Philosophy, Illinois Wesleyan University, 205 Beecher Street, Bloomington, IL, 61701. Email: tblancha@iwu.edu

intuitive judgments correspond to BOR in favoring less flexible hypotheses, whether making estimates of probability or evaluating the quality of explanations.

According to Bayes' theorem, the relative credibility of two hypotheses H_1 and H_2 in light of data D, $\frac{P(H_1/D)}{P(H_2/D)}$, equals

$$\frac{P(H_1)P(D/H_1)}{P(H_2)P(D/H_2)}$$

The first ratio in this expression measures the extent to which H_1 is initially more plausible than H_2, while the second—the *likelihood ratio* of H_1 and H_2—measures how well H_1 predicts D compared to H_2. The idea behind the BOR is that by its very nature a hypothesis that is more complex (in the sense that it contains more free parameters or more possible parameter values) tends to predict the actual data less well than a simpler hypothesis. The reason is that a complex hypothesis is more *flexible*: By adjusting or "tuning" the parameters in the right way, the hypothesis can be made to accommodate a wide range of possible data. The flipside, however, is that for many parameter settings, the hypothesis fits the actual data very poorly, so that, assuming a relatively uniform probability distribution over the parameter space, overall the probability of the data under the hypothesis will be relatively low. Hence, as long as H_1 fits the data relatively well, $P(D/H_2)$ will be lower than $P(D/H_1)$, so that unless H_2 is initially much more plausible than H_1, Bayesian inference will tend to favor H_1 over H_2.

To illustrate, suppose that a coin is tossed 10 times and comes up heads 4 times. One hypothesis is that the coin is fair (H_1), which gives a probability of .21 to the data. Another, more flexible hypothesis H_2 says that the probability of heads is n/10, where n is equally likely to be any natural number between 1 and 8. Here, the probability of heads is a free parameter, that is, a parameter that can take a variety of possible settings. This parameter can be "adjusted" to fit a variety of possible sequences, and in fact can be adjusted to fit the actual sequence better than H_1: stipulating $n = 4$ yields a probability of .25 for the observed sequence. But other possible settings of the parameter yield a poor fit with the observed sequence. Since the probability distribution over the space of possible settings of the free parameter is uniform (i.e., each parameter setting is equally likely), overall H_2 fits the actual data more poorly than H_1. Specifically, H_2 assigns a probability of only .11 to the observed sequence.[1] If the two hypotheses are a priori equally likely, the evidence favors the less flexible hypothesis H_1.

Note that the BOR is sensitive not only to the *number* of free parameters in a hypothesis, but also to the *size* of the parameter space, that is, the number of possible values that a parameter may take. Thus, a third hypothesis H_3, where n is allowed to be any natural number between 1 and 10, yields an even poorer fit with the data, as the additional values of the free parameter assign a particularly low probability to the actual sequence of tosses. This is one of the main differences between the BOR and other criteria for hypothesis selection such as AIC (Akaike, 1974; Forster & Sober, 1994) and BIC

(Schwarz, 1978), which penalize hypotheses solely as a function of the number of free parameters they contain, without an additional penalty for the number of values that those parameters can assume.

Philosophers of science have shown that the considerations related to the BOR plausibly explain various aspects of scientific inference (Henderson et al., 2010), including major historical episodes of scientific theory change such as the dispute between Ptolemaism and Copernicanism (Henderson, 2014; Myrvold, 2003). Moreover, there is considerable psychological evidence that people engage in Bayesian inference (see Griffiths, Tenenbaum, & Kemp, 2012 for an overview), although the most interesting conditions for testing the BOR—cases in which parameterized hypotheses vary in flexibility—have not been investigated. There is also evidence that people's inferences are guided by considerations of simplicity (Bonawitz & Lombrozo, 2012; Lombrozo, 2007, 2016; Pacer & Lombrozo, in press; Read & Marcus-Newhall, 1993), but this work has evaluated a different measure of simplicity (the number of assumptions or unexplained causes invoked in an explanation), and evidence that people respond to this measure has taken the form of *departures* from probabilistic inference.

We report the results of two experiments investigating whether people penalize more flexible hypotheses in accordance with the BOR. Our experiments assess hypothesis evaluation across two kinds of judgments: *probability* and *explanation*. We assess the former by having participants indicate which of two hypotheses they think is more "likely" in light of some observations. We assess the latter by having participants indicate which of two hypotheses they think is a "better explanation" for those observations. We deliberately did not provide further guidance on what constitutes a better explanation. This allowed us to avoid commitment to a specific theory of explanation, such as the causal or unification account (for reviews, see Lombrozo, 2011, 2012; Woodward, 2017).

Our study included both probabilistic and explanatory judgments because the two have been shown to diverge in the context of hypothesis choice (Douven & Schupbach, 2015). In particular, it could be that explanatory judgments are more sensitive to likelihoods than to priors (Douven & Schupbach, 2015; Pacer, Williams, Chen, Lombrozo, & Griffiths, 2013), resulting in a greater penalty for flexible hypotheses when the hypotheses are evaluated as explanations for a set of observations.

2. Experiment 1

The main goal of Experiment 1 was to examine whether people's probabilistic and explanatory judgments conform to the BOR in penalizing more flexible hypotheses. To do so, we asked participants which of two hypotheses, H_1 and H_2, was more credible in light of the presented data. We assigned participants to one of three conditions: the degree of flexibility of H_1 was kept constant, but the number of parameters included in H_2 and the size of the relevant parameter space varied across the three conditions.

2.1. Method

2.1.1. Participants

A total of 178 participants (46% women, mean age 34, range 18–72) were recruited online on Amazon Mechanical Turk and paid $0.50 for their participation. An additional 68 participants were excluded for failing a comprehension check. In all experiments, participation was restricted to users with an IP address within the United States and an approval rating of at least 95% based on at least 50 previous tasks.

2.1.2. Materials, design, and procedure

Participants were placed in the role of a scientist on a fictional planet studying two "almost indistinguishable" and "equally common" frog-like species, the "velmos" and the "zorgits," both of which commonly have red spots on their backs. Participants read information about the frequency of red spots for each species.

For velmos, the information presented to participants was always the same:

> Around 50% of [velmos] have no red spots, 25% of them have one red spot, and 25% have two red spots. Thus if you observe 100 velmos, a representative sample would include 50 with no red spots, 25 with one red spot, and 25 with two red spots. In addition, velmos inherit their number of red spots from their mother. So if a female velmo has no spots, her offspring will have no spots; if she has one spot, her offspring will have one spot; and if she has two spots, her offspring will have two spots.

For zorgits, on the other hand, the information varied across conditions: the No Parameter condition, the Medium Parameter condition, and the Large Parameter condition ("No," "Medium," and "Large" for short). In the No condition, participants were told that zorgits "always have two red spots on their back." In the Medium condition, participants read the following:

> Zorgits can have anywhere between 1 and 4 red spots on their back, and each of these possibilities is equally likely. Thus if you observe 100 zorgits, a representative sample would have 25 with one red spot, 25 with two red spots, 25 with three red spots, and 25 with four red spots.

In the Large condition, participants were told that zorgits can have anywhere between 1 and 100 spots on their back, so that:

> if you observe 1000 zorgits, a representative sample would have 10 of them with one red spot, 10 of them with two red spots, 10 of them with three red spots, and so on up to 100 red spots.

In the Medium and Large conditions, participants also read that zorgits inherit the number of red spots on their back from their mother.

Participants were then told that their research assistants just discovered a small family of frog-like animals consisting of a female and two of its babies, each with two red spots. After reading this information, participants were asked to choose between two competing hypotheses: that the family is a family of velmos (H_1), or that it is a family of zorgits (H_2). They evaluated both which was more likely and which was a better explanation (in counterbalanced order), using a forced choice followed by a 3-point scale (see Table 1). This allowed us to measure participants' estimates of the comparative probability/explanatoriness of H_1 and H_2 on a 6-point scale, from 1 (H_1 judged much more likely or explanatory than H_2) to 6 (H_2 judged much more likely or explanatory than H_1). We call these *comparative strength* ratings.

The descriptions across the three conditions were designed to vary the presence of a free parameter in hypothesis H_2 (No vs. Medium and Large) as well as the number of values the parameter could take on (Medium vs. Large). Specifically, whereas the number of spots on the zorgit mother's back is fixed in No, this parameter becomes free in Medium and Large, in the sense that it can take various different values yielding different probabilities for the data. And the number of possible values of this parameter increases from 4 in Medium to 100 in Large. (In both conditions, the probability distribution over the parameter space is uniform.) In contrast, the number of free parameters in H_1 (and the size of the associated parameter space) remains constant over the three conditions. Specifically, in all conditions, H_1 has one free parameter—the number of spots on the velmo mother's back—with possible values 0, 1, and 2 (whose respective probabilities are ½, ¼, and ¼).

As H_2 becomes more flexible across the three conditions (either by having more free parameters or by having parameters with more possible values), it becomes increasingly penalized by the BOR. In contrast, H_1's flexibility (and hence its likelihood) remains constant. As a result, the likelihood ratio $LR(H_1, H_2)$ increases across conditions, and since the prior probabilities of H_1 and H_2 are the same (zorgits and velmos are equally common), the posterior ratio of the two hypotheses is the same as their likelihood ratio (see Table 2).[2]

Table 1
Questions in Experiment 1, as a function of judgment (probability vs. explanation)

	Probability	Explanation
Q1[a]	Given that all three members of the family have two spots on their back, which of the following do you think is more likely? (A) This is a family of zorgits. (B) This is a family of velmos.	In your opinion, which of the following is a better explanation for the fact that all three members of this family have two red spots on their back? (A) This is a family of zorgits. (B) This is a family of velmos.
Q2	How much more likely do you think the option you chose is than the alternative? (A) Slightly more likely (B) Moderately more likely (C) Much more likely	How much better do you think the explanation you chose is than the alternative? (A) Slightly better (B) Moderately better (C) Much better

Note. [a]Answers for Question 1 were presented in random order.

Table 2
Likelihood and likelihood ratio of H_1 and H_2 as a function of parameter condition in Experiment 1

	No		Medium		Large	
	H1	H2	H1	H2	H1	H2
Likelihood	0.25	1	0.25	0.25	0.25	0.01
LR (H1, H2)	0.25		1		25	

Thus, if people are sensitive to the flexibility of a hypothesis when assessing hypothesis strength, we should expect H_2's comparative strength to decrease across parameter conditions. In particular, participants should ascribe less comparative strength to H_2 in Medium than in No, given the introduction of a free parameter, and in Large than in Medium, given the additional values that the free parameter may assume.

At the end of the task, participants were asked why they chose the hypothesis they did, and given the option to type a few sentences in a text box. We call this the *justification* of their hypothesis choice.

2.2. Results and discussion

2.2.1. Comparative strength

A 3 parameter (no, medium, large) × 2 judgment (explanation, probability) ANOVA on comparative strength ratings revealed a significant main effect of parameter, $F(2, 178) = 48.90$, $p < .001$, $\eta_p^2 = .363$ (see Fig. 1). Post hoc independent samples *t*-tests revealed that the comparative strength of H_1 versus H_2 increased significantly across parameter conditions. Strength judgments favored H_1 significantly more in Large ($M = 2.55$) than in Medium ($M = 3.95$), $t(113) = -6.33$, $p < .001$, or No ($M = 4.62$),

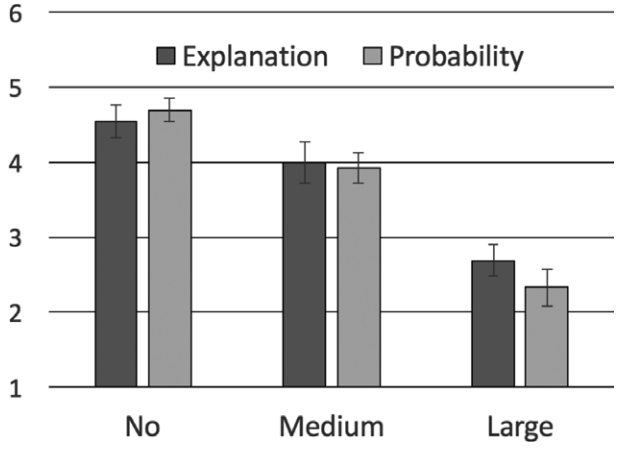

Fig. 1. The effect of parameter condition on judgments of comparative probabilistic and explanatory strength in Experiment 1.

$t(121)=-9.47$, $p < .001$. Judgments also favored H_1 significantly more in Medium than in No, $t(116) = -3.17$, $p = .002$. There was no significant main effect of judgment ($p = .677$), nor a significant interaction ($p = .485$). This suggests that flexibility has an effect on both probabilistic and explanatory judgments, and that this effect is not moderated by the nature of the judgment.

2.2.2. Justification

Two coders classified participants' justifications into one or more of five coding categories (see Table 3). Disagreements between coders were resolved through discussion (average Cohen's $\kappa = .64$, $p < .001$ for each category). Excluding the minority of participants who misunderstood the task from analysis does not change the conclusions presented above.

It is noteworthy that while most comments made reference to probability, only a small number of comments (N = 11) explicitly appealed to considerations related to flexibility.

Overall, the results of Experiment 1 provide evidence that participants are sensitive to the form of simplicity envisioned by the BOR and tend to penalize flexibility when assessing both the probability and explanatory strength of a hypothesis, including when the increase in flexibility is due to an increase in parameter space size rather than number of parameters.

3. Experiment 2

Our goals in Experiment 2 were to address two possibilities left open by Experiment 1. First, for the scenario in Experiment 1, the respective likelihoods (and posterior probabilities) of the two hypotheses were relatively easy to compute. Thus, Experiment 1 left open the possibility that people are only sensitive to Bayesian considerations of flexibility when they can effectively "do the math," and not when they must rely on more intuitive

Table 3
Coding categories used to classify justifications for hypothesis choices in Experiments 1 and 2, along with the proportion of participants who produced each justification type

Coding Category	Criterion	Proportion (Exp. 1)	Proportion (Exp. 2)
Probability	Made reference to the probability of a hypothesis (either likelihood or posterior)	.61	.48
Flexibility	Referred to features of the stimuli related to flexibility	.06	.11
Precise	Mentioned precise numbers	.31	.04
Other	Offered some justification for hypothesis choice other than flexibility or probability	.28	.37
Misunderstood	Response suggested participant misunderstood at least one aspect of scenario	.08	.05

Note. Proportions do not sum to 1, as a single justification could fall into more than one category.

assessments. Second, Experiment 1 did not test one particularly important consequence of the BOR, namely that a hypothesis can be penalized for flexibility even when the parameters of the hypothesis can be "tuned" to fit the data *better* than a comparatively less flexible hypothesis (recall the coin-flipping example from the introduction). Thus, the results of Experiment 1 were consistent with an alternative hypothesis on which probabilistic and explanatory judgments are sensitive to the likelihood of the hypothesis on the parameter setting that best fits the data, and only penalize more flexible hypotheses when the relevant likelihoods are equal, in contradiction with the BOR's verdicts. The goal of Experiment 2 was to address these two possibilities.

3.1. Method

3.1.1. Participants

One hundred and eighty participants (44% women, mean age 34, range 19–83) were recruited online on Amazon Mechanical Turk and paid $0.50 for their participation. An additional 91 participants were excluded for failing a comprehension check.

3.1.2. Materials, design, and procedure

The design and procedure of Experiment 2 were the same as those of Experiment 1, but the scenario differed. Participants read a story involving a family (the Millers) who every year prepare candy bowls for trick-or-treaters on Halloween. Participants were told that every year, Mr. and Mrs. Miller each fill a pumpkin with bags of Skittles and M&Ms. In all conditions, participants were told that "Mrs. Miller thinks that M&Ms are a bit tastier than Skittles, so she always fills her pumpkin with 160 bags of M&Ms and 140 bags of Skittles." By contrast, the information provided about Mr. Miller varied across the three parameter conditions. In the No condition, participants read:

> because he is more frugal than Mrs. Miller, Mr. Miller only places 200 bags of candies in his pumpkin. And because he thinks that M&Ms and Skittles taste equally good, he always puts in exactly 100 bags of M&Ms and 100 bags of Skittles.

In the Medium condition, participants read:

> Because Mrs. Miller is so predictable, he likes to be unpredictable. So every year he has his computer generate two random numbers, each between 91 and 100. The first number determines how many bags of M&Ms he puts in his pumpkin, and the second number determines how many bags of Skittles he puts in his pumpkin.

In the Large condition, the text was the same as in Medium, except that the random generator outputs two numbers between 1 and 100, so that in a given year Mr. Miller's pumpkin can contain anywhere between 1 and 100 of bags Skittles and 1 and 100 bags of M&Ms.

Participants were then told that on one Halloween night, a group of kids arrives at the Miller's door; their son opens the door and randomly selects one of the two pumpkins, from which the kids are allowed to pick bags of candies at random. In total, the kids get 99 bags of M&Ms and 99 bags of Skittles.

Participants were then asked to choose between two hypotheses—that "the kids picked their candies from Mrs. Miller's pumpkin" (H_1) or that "the kids picked their candies from Mr. Miller's pumpkin" (H_2), following the same procedure as Experiment 1 (see Table 1). Participants were also asked to provide a justification for their hypothesis choice.

As in Experiment 1, the flexibility of H_1 remains constant across all three parameter conditions, whereas the flexibility of H_2 increases across the parameter conditions along two dimensions. First, H_2 contains two extra free parameters in the Medium and Large conditions compared to the No condition, namely the two outputs of the random number generator. Second, the number of possible values for each parameter increases from Medium to Large, so that H_2 is increasingly penalized by the BOR (see Table 4).[3] But there are two salient differences from Experiment 1. First, the likelihoods of the two hypotheses are much more difficult to compute than in Experiment 1: it is unlikely that participants could effectively "do the math." Second, in the Medium and Large condition, the extra flexibility of H_2 means that it can be tuned to fit the data *better* than H_1. Indeed, together with the auxiliary hypothesis that the random generator output 99 twice, the likelihood of H_2 is 1 in both Medium and Large.

3.2. Results and discussion

3.2.1. Comparative strength

A 3 parameter (no, medium, large) × 2 judgment (explanation, probability) ANOVA once again revealed a significant main effect of parameter condition on comparative strength ratings, $F(2, 180) = 26.95$, $p < .001$, $\eta_p^2 = .236$ (see Fig. 2). Post hoc independent samples t-tests revealed that as in Experiment 1, judgments of comparative strength favored H_1 significantly more in Large ($M = 2.51$) than in Medium ($M = 3.39$), $t(108) = -3.12$, $p = .002$, and No ($M = 4.34$), $t(117) = -7.38$, $p < .001$). In addition, judgments of comparative strength favored H_1 significantly more in Medium than in No, $t(129) = -3.92$, $p < .001$.

The ANOVA on comparative strength ratings revealed no significant main effect of judgment ($p = .321$). However, and by contrast to Experiment 1, there was a

Table 4
Likelihood and Likelihood ratios of H_1 and H2 as a function of parameter condition in Experiment 2

	No		Medium		Large	
	H1	H2	H1	H2	H1	H2
Likelihood	0.026	0.5	0.026	0.025	0.026	0.00025
LR(H1, H2)	0.052		1.04		104	

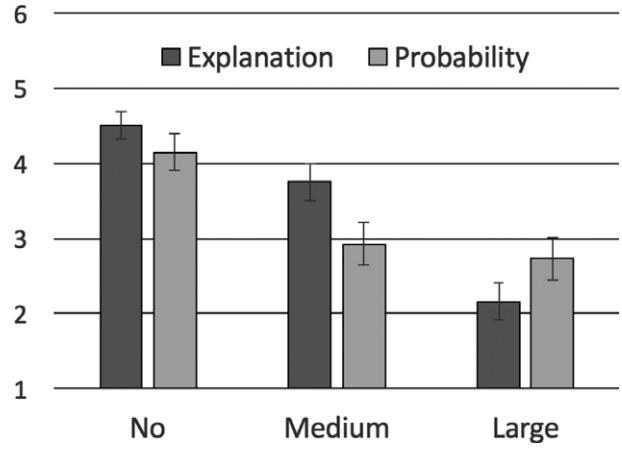

Fig. 2. The effect of parameter condition on judgments of comparative probabilistic and explanatory strength in Experiment 2.

significant interaction between parameter and judgment, $F(2, 180) = 3.54$, $p = .031$. Compared to probability judgments, explanatory judgments penalized the more flexible hypothesis less in the Medium condition, but more in the Large condition. This raises the possibility that the effect of flexibility on hypothesis assessment differs significantly for explanatory and probabilistic judgments—a possibility worth taking seriously in light of the fact that systematic deviations between probabilistic and explanatory judgments have already been documented (Douven & Schupbach, 2015). Nevertheless, as far as we can see there is no plausible theoretical explanation for why explanatory and probabilistic judgments should differ in the way observed here: it is hard to see why explanatory judgments should be *less* sensitive to flexibility than probabilistic judgments when the hypothesis is somewhat flexible (as in Medium), but become *more* sensitive to flexibility than probability judgments when the hypothesis is very flexible (as in Large). Moreover, the effect was not observed in Experiment 1. We are therefore hesitant to draw strong conclusions and think that the issue requires further investigation.

3.2.2. *Justification*

Participants' justifications were coded as in Experiment 1 (average Cohen's $\kappa = .65$, $p < .001$, for all categories; see Table 3). Only a small percentage (5%) of participants' justifications were coded as "misunderstood," and excluding these participants from analysis did not change the results reported above (except that the parameter × judgment interaction became marginally significant, $F(2, 171) = 2.9$, $p = .058$). By comparison to Experiment 1, a significantly smaller percentage (3.9%) of participants mentioned precise numbers in their comments, $\chi^2(1, N = 358) = 40.89$, $p < .001$, confirming that precise probabilities were much more difficult to compute in Experiment 2 than in Experiment 1. As in Experiment 1, only a small number of comments (19) appealed to considerations of

flexibility, suggesting that such considerations do not play a reliable role in explicit reasoning.

Overall, the results of Experiment 2 provided further confirmation that probabilistic and explanatory judgments are sensitive to the penalty for flexibility induced by the BOR. In particular, probabilistic and explanatory judgments appropriately penalize a hypothesis for flexibility even when the flexibility of the hypothesis means that it can be "tuned" to fit the data better than a less flexible hypothesis, and when the relevant probabilities are difficult to compute. Moreover, the explanatory and probabilistic judgments both show this penalty for flexibility.

4. General discussion

Our two experiments provide evidence that people's intuitive judgments follow the prescriptions of BOR, whether making estimates of the probability of a hypothesis or evaluating how well the hypothesis explains the data. In particular, people's judgments are consistent with the two most distinctive characteristics of BOR: They penalize hypotheses as a function of their *flexibility* (which is determined not only by the number of free parameters but also by the size of the parameter space), and they penalize those hypotheses even when their parameters can be "tuned" to fit the data better than comparatively simpler hypotheses.

Our results go beyond previous demonstrations of an intuitive preference for simpler hypotheses. Prior work has shown that people do seem to favor explanations that are simpler in that they involve fewer independent assumptions (Lombrozo, 2007, 2016; Read & Marcus-Newhall, 1993) or "root causes" (Pacer & Lombrozo, in press), and that this could reflect a preference built into the prior probabilities assigned to hypotheses in a given domain. However, these studies do not test the idea that a preference for simplicity results from the mechanics of Bayesian inference itself; in fact, the preference for "root simplicity" manifests itself as a preference for the root-simpler explanation when this choice is not warranted by an application of Bayes' rule using the probabilistic information provided in the task. Our results instead show that people's intuitive judgments are sensitive to a different form of simplicity (inflexibility), in a way that is perfectly consistent with the verdicts of the Bayesian account of inference.

Our results also go beyond previous demonstrations that people's preference for simplicity can in certain cases be explained along Bayesian lines. For instance, Tenenbaum and Griffiths (2003) show that people's causal inferences exhibit a preference for simpler hypotheses, where this preference can be readily explained in Bayesian terms. But the version of Occam's razor they are concerned with does not penalize hypotheses for their flexibility (i.e., their ability to accommodate a wide range of possible data); instead, it penalizes hypotheses that posit more causes than are necessary to explain the actual data. In addition, there is evidence that people follow a "size principle" when engaging in generalization and rule-learning tasks (e.g., Tenenbaum & Griffiths, 2001): among all the hypotheses consistent with the observed stimuli, people tend to prefer

more specific hypotheses to more general alternatives. For instance, if the observed stimuli consist of Siamese cats who have a certain property P, people tend to prefer the hypothesis "All Siamese cats have P" to the more general hypothesis "All cats have P." As Tenenbaum and Griffiths argue, this behavior can be explained in Bayesian terms, as more specific hypotheses tend to have higher likelihoods on the observed data than more general hypotheses. Yet, here again, this penalty is not a penalty for *flexibility*. More general hypotheses (in the sense of "generality" at work in the size principle) are not more flexible than less general ones: They do not contain more free parameters, and thus cannot be made to accommodate a wider range of possible data. Finally, there is also evidence that people's priors tend to favor hypotheses that posit a small number of strong causes (Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2008; Powell et al., 2016). Yet, in this case, the preference for simplicity is built into the priors (see also Lombrozo, 2007); our results instead suggest that a preference for simpler hypotheses can emerge as a consequence of the Bayesian preference for hypotheses with higher likelihoods.

A potential limitation of the present studies is that in both experiments, evaluating the relative flexibility of the competing hypotheses was relatively straightforward. The rival hypotheses had similar forms, posited the same kind of mechanism to explain the data, and were explicitly presented with the information required to evaluate both the size of the relevant parameter spaces and the prior probability distributions over them. In other contexts, comparative judgments of flexibility may be more complex or indirect. For instance, when competing hypotheses posit very different causal mechanisms or entities to explain the data, participants may rely on more heuristic guides to relative flexibility. The literature on model selection in cognitive psychology illustrates some of these challenges: When comparing competing theories of a psychological phenomenon—such as theories of information integration (Myung & Pitt, 1997) or theories of decision-making under uncertainty (Glöckner & Pachur, 2012)—in a way that takes flexibility into account, counting the number of parameters contained in each theory often requires turning each of them into a precise mathematical model. Whether and how judgments of probability and explanation quality are sensitive to BOR-driven considerations in contexts where flexibility assessments are less straightforward is an important topic for further investigation.

In our view, the main interest of our results is their relevance to an important issue about the status of explanatory considerations in reasoning. There is considerable evidence that explanatory considerations—and especially considerations of simplicity—play a central role in learning and inference (Lombrozo, 2016). Why is this the case? A similar normative question arises in the philosophy of science, where one popular answer is that the rational bearing of explanatory virtues (and in particular simplicity) on inference is a straightforward consequence of Bayesianism itself (Henderson, 2014; Myrvold, 2003). Our results suggest that this may be partially true at the descriptive level as well. That is, people's preference for simpler hypotheses may in part be a natural consequence of the fact that their judgments approximate Bayesian inference—although it is unlikely

that all effects of explanatory considerations in reasoning can be explained in this way (Douven & Schupbach, 2015).

Acknowledgments

The authors thank the John Templeton Foundation's Varieties of Understanding project for funding this research.

Notes

1. Note that a definite likelihood for a hypothesis containing free parameters can be calculated only given a prior probability distribution over the space of possible parameter settings. In some cases it may be unclear what the right probability distribution over the parameter space actually is. However, in all the cases considered in this paper, the probability distribution over the parameter space is explicitly stipulated.
2. It is worth asking how H_1 and H_2 fare under other criteria for hypothesis selection, such as AIC and BIC, which also penalize hypotheses as a function of their flexibility. Two remarks are in order here. First, AIC and BIC were originally developed as solutions to a very specific hypothesis selection problem—the "curve-fitting" problem. Because the problem presented to participants differs in important ways from the curve-fitting problem, applying AIC or BIC to the case of H_1 and H_2 is not entirely straightforward. (Note also that while BIC, like the BOR, is concerned with likelihood—both embody the idea that *ceteris paribus* more flexible hypotheses make the actual data less probable—AIC is not concerned with likelihood at all. Instead, AIC embodies the idea that more flexible hypotheses are less *predictively accurate*, that is, fare worse at correctly predicting new data drawn from the same underlying distribution.) Second, as we noted in the Introduction, AIC and BIC penalize hypotheses solely as a function of the number of free parameters that they contain, and they are insensitive to the size of the parameter space. (In contexts in which AIC and BIC are usually applied, such as the curve-fitting problem, issues of parameter space size do not arise, as all the parameters under consideration can take continuously many values.) This means that insofar as they can be coherently applied to the case at hand, AIC and BIC penalize H_2 more in Medium and Large than in No, but do not penalize it more in Large than Medium, since H_2 has the same number of free parameters in both conditions. Thus, AIC and BIC do not predict a larger preference for H_1 in Large than in Medium.
3. Because H_2 has the same number of free parameters in both Medium and Large, AIC and BIC do not penalize the hypothesis more in the latter than in the former (insofar as they can be coherently applied to the case), mirroring Experiment 1.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Baker, A. (2013). Simplicity. In E. N. Zalta(Ed.), *The Stanford Encyclopedia of Philosophy (Fall 2013 Edition)*, Available at http://plato.stanford.edu/archives/fall2013/entries/simplicity/. Accessed May 16, 2017.

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

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

Forster, M., & Sober, E. (1994). How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. *British Journal for the Philosophy of Science*, *45*, 1–35.

Glöckner, A., & Pachur, T. (2012). Cognitive models of risky choice: Parameter stability and predictive accuracy of prospect theory. *Cognition*, *123*, 21–32.

Griffiths, T. L., Tenenbaum, J. B., & Kemp, C. (2012). Bayesian Inference. In K. J. Holyoak, & R. G. Morrison (Eds.), *Oxford handbook of thinking and reasoning* (pp. 22–35). New York: Oxford University Press.

Henderson, L. (2014). Bayesianism and inference to the best explanation. *British Journal for the Philosophy of Science*, *65*, 687–715.

Henderson, L., Goodman, N. D., Tenenbaum, J. B., & Woodward, J. (2010). The structure and dynamics of scientific theories: A hierarchical Bayesian perspective. *Philosophy of Science*, *77*, 172–200.

Jefferys, W. H., & Berger, J. O. (1992). Ockham's razor and bayesian analysis. *American Scientist*, *80*, 64–72.

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

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

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

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

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review*, *115*, 955–984.

MacKay, D. J. C. (2003). *Information theory, inference, and learning algorithms*. Cambridge, UK: Cambridge University Press.

Myrvold, W. (2003). A Bayesian account of the virtue of unification. *Philosophy of Science*, *70*, 399–423.

Myung, J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, *4*, 79–95.

Pacer, M., & Lombrozo, T. (in press). Ockham's razor cuts to the root: Simplicity in explanation choice. *Journal of Experimental Psychology: General*.

Pacer, M., Williams, J., Chen, X., Lombrozo, T., & Griffiths, T. L. (2013). Evaluating computational models of explanation using human judgments. In A. Nicholson & P. Smyth (Eds.), *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence* (pp. 498–507). Corvallis, OR: AUAI Press

Powell, D., Merrick, M. A., Lu, H., & Holyoak, K. J. (2016). Causal competition based on generic priors. *Cognitive Psychology*, *86*, 62–86.

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.

Rosenkrantz, R. D. (1977). *Inference, method, and decision: Towards a Bayesian philosophy of science*. Dordrecht, the Netherlands: Synthese Library.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–465.

Sober, E. (2015). *Ockham's razors*. Cambridge, UK: Cambridge University Press.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*, 629–640.

Tenenbaum, J. B., & Griffiths, T. L. (2003). Theory-based causal inference. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems 15* (pp. 67–74). Cambridge, MA: MIT Press.

Woodward, J. (2017). Scientific Explanation. In E.N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2017 edition). Available at https://plato.stanford.edu/archives/spr2017/entries/scientific-explanation/. Accessed May 16, 2017.