Simplicity in Explanation

Tania Lombrozo (lombrozo@wjh.harvard.edu)
Department of Psychology, Harvard University
33 Kirkland St., Cambridge, MA 02144

J. Jane Rutstein (jessica.rutstein@tufts.edu)
Department of Philosophy, Tufts University
Medford, MA 02155

Abstract

In this paper we explore the role of simplicity in choosing between competing explanations, and in particular how a preference for simplicity is integrated with information about the probability of particular explanations. In Experiment 1 we establish that all else being equal, people prefer explanations that are simpler in the sense of invoking fewer causes. Experiment 2 finds that people require disproportionate evidence in favor of a complex explanation before they will choose it over a simpler alternative. Experiment 3 suggests that this bias is not driven by assumptions about the probabilistic dependence of causes. Finally, Experiment 4 replicates the basic findings with a more ecologically valid computer task. We also find that participants who prefer a simpler but less probable explanation overestimate the frequency of events that would make the simpler explanation more probable. We conclude by suggesting that people believe simpler explanations are more likely to be true in virtue of being simple.

Introduction

Explaining the world around us is a fundamental part of everyday life. We wonder why objects have the properties they do, why people act in particular ways, and why things do or don't happen. But more often than not, explanations are vastly underconstrained by our knowledge and the available data. When more than one explanation is possible, how do we choose between them? A plausible constraint on competing explanations, often attributed to William of Occam, is simplicity. Here we explore whether people in fact prefer simpler explanations, and if so how they balance a preference for simplicity with the desire to maximize other virtues of explanation, like their probability of being true. We show that people do prefer simpler explanations, even when they are less probable than more complex alternatives. We also show that this preference can lead to systematic distortions in the perceived frequency of events.

A Metric for Simplicity

While simplicity is commonly invoked, it is notoriously difficult to formalize and justify. Several recently proposed approaches, like the Akaike Information Criterion (AIC) (e.g. Sober, forthcoming), Bayesian Occam's razor (e.g. Jeffreys & Berger, 1992), Minimum Description Length (MDL) and Kolmogorov Complexity (e.g. Chater &

Vitanyi, 2003), nonetheless succeed in precisely specifying a metric for simplicity in the language of statistics and computer science. What's more, these metrics can be motivated on principled grounds. The AIC warrants a preference for simplicity by showing that simpler explanations are more likely to generalize. Similarly, Bayesian Occam's Razor shows that a simpler explanation will have a higher posterior probability. From a bottom-up perspective, considerations of processing constraints make measures like MDL and Kolmogorov complexity attractive.

While compelling, formal measures of simplicity are generally formulated over well-defined problems like line fitting, which bear little resemblance to the complex inductive leaps that characterize everyday explanatory judgments. For this reason we looked to the history of science for a more psychologically plausible metric. In the *Principia*, Newton wrote that "we are to admit no more causes of natural things than such as are both true and sufficient to explain their appearances" (1686). This maxim, similar to Occam's statement that entities should not be multiplied beyond necessity, suggests that explanations invoking fewer causes are to be preferred. We thus chose to quantify simplicity in terms of number of causes, where explanations involving fewer causes are simpler.

Simplicity and Probability

Newton's endorsement of explanations with fewer causes was likely grounded in metaphysical assumptions. After the quote above, he went on to suggest that "nature is pleased with simplicity and affects not the pomp of superfluous causes." If nature is in fact simple, then simple explanations are more likely to be true.

In the first experiments reported below we explore whether people prefer explanations involving fewer causes, but also if this preference is motivated by the belief that simpler explanations are more likely to be true. We examine the relationship between simplicity and probability both directly and indirectly. As a direct test, we look at people's justifications for choosing a simpler explanation. More indirectly, we look at whether people switch their preference from a simpler to a more complex explanation when provided with evidence that the more complex explanation is more likely.

Seeing how people balance the competing explanatory virtues of simplicity and probability can help distinguish

two possible hypotheses about the nature of a preference for simplicity. According to what we call the *probabilistic metric* hypothesis, people prefer simpler explanations, but represent this preference in terms of probability. That is, simpler explanations are believed to be more likely to be true in virtue of being simple. While choosing between competing explanations involves deciding which is most likely to be true, simpler explanations gain a probabilistic boost just for being simple. A second hypothesis is the *trade-off* hypothesis, according to which simplicity and probability are independent virtues of explanation that must be integrated according to some weighting function. The *probabilistic metric* hypothesis differs from the *trade-off* hypothesis in that the former claims the preference for simpler explanation is expressed in terms of probability, whereas the latter assumes that simplicity and probability trade-off in a way that is not commensurate.

In the final experiment we go on to explore the consequences of a tendency to favor simpler explanations. Specifically, does the preference for simpler explanations distort our perception of probability? If so, we would expect people's preferred explanations to influence their frequency judgments.

Experiments

All experiments we report involve a simple task adapted from Lagnado (1994) in which participants are asked to choose between one and two diseases to account for some symptoms. By varying the prevalence of the diseases we were able to manipulate the relative probability of the simpler, one-disease explanation to the more complex, two-disease explanation.

Experiment 1: Explanatory Virtues

Before examining how people integrate information about simplicity and probability in explanation, we wanted to confirm that simplicity and probability are indeed virtues of explanation.

Methods Twenty-four Boston-area undergraduate and summer school students completed a questionnaire in one of two conditions: the *simplicity* condition and the *probability* condition.

In the *simplicity* condition, participants read the following:

> There is a population of 750 aliens that lives on planet Zorg. You are a doctor trying to understand an alien's medical problem. The alien, Treda, has two symptoms: Treda's **minttels are sore** and Treda has developed **purple spots**.
>
> **Tritchet's syndrome** always causes both **sore minttels** and **purple spots**.
>
> **Morad's disease** always causes **sore minttels**, but the disease never causes **purple spots**.
>
> When an alien has a **Humel infection**, that alien will always develop **purple spots**, but the infection will never cause **sore minttels**.

Nothing else is known to cause an alien's **minttels to be sore** or the development of **purple spots**.

They were then asked to choose the *most satisfying* explanation for Treda's symptoms among a list of possibilities that included every disease individually and every pairwise combination of diseases. Choosing Tritchet's syndrome would be the simplest option; choosing Morad's and a Humel infection is a more complex alternative.

In the *probability* condition, the cover story was similar, but participants were asked to choose between two diseases that each accounted for both symptoms. However, one disease was said to be present in about 50 of the aliens on Zorg, while the other was present in about 73 aliens on Zorg, making the latter choice the more probable option.

After choosing an explanation, participants were also asked to explain their reasoning. The names of the diseases were counterbalanced and we used three different sets of symptoms.

Results and Conclusions In the *simplicity* condition, 100% of participants chose the simpler explanation. They justified this choice about equally often by appeal to simplicity and probability: 50% explicitly said they chose it because it was simpler, while 42% said they thought it was more likely for the alien to have one disease than two. In the *probability* condition, 92% of participants chose the more probable explanation. All participants justified this choice by appeal to probability.

Experiment 2: Simplicity Versus Probability

Having established that people do prefer both simpler and more probable explanations, we went on to see how these virtues of explanations are traded off. To do so we had participants choose explanations in cases where the simplest was not the most likely to be true.

Methods One-hundred-thirty-seven Boston-area summer school and undergraduate students participated by completing a questionnaire. The questionnaire was like the *simplicity* condition from Experiment 1, but participants were additionally given information about the prevalence of each disease in the population. For example, one questionnaire read:

> There is a population of 750 aliens that lives on planet Zorg. You are a doctor trying to understand an alien's medical problem. The alien, Treda, has two symptoms: Treda's **minttels are sore** and Treda has developed **purple spots**.
>
> **Tritchet's syndrome** always causes both **sore minttels** and **purple spots**. **Tritchet's syndrome** is present in about 50 aliens on Zorg.
>
> **Morad's disease** always causes **sore minttels**, but the disease never causes **purple spots**. **Morad's disease** is present in about 225 of the aliens on Zorg.
>
> When an alien has a **Humel infection**, that alien will always develop **purple spots**, but the infection will never cause **sore minttels**. You know that Humel Infections are present in about 210 of the aliens on Zorg.

As in Experiment 1, they were then asked to choose the most satisfying explanation and selected among six options, which included each disease individually and every pairwise combination. On a second page of the questionnaire they were asked to justify their choice, and also to complete a math problem. The math problem required participants to compute the joint probability of winning at two slot machines and compare this to the probability of winning at a different machine. We included this problem to see whether participants knew how to compute joint probabilities.

We varied the prevalence of the diseases to manipulate the relative probability of having the single disease causing both symptoms (D_1) to having both of the other diseases (D_2&D_3). Table 1 indicates the 8 sets of values we used, along with the corresponding probability ratios, which were computed on the assumption that the diseases are probabilistically independent. There were 14 to 18 participants per condition.

Table 1: Disease prevalence for each frequency condition.

D_1	D_2	D_3	$P(D_1):P(D_2$&$D_3)$
50	50	50	15:1
50	197	190	1:1
50	195	214	9:10
50	225	210	4:5
50	250	220	2:3
50	268	280	1:2
50	330	340	1:3
50	610	620	1:10

Explanation justifications were coded into one of three categories: simplicity, probability, and other. Justifications were coded as 'simplicity' if (1) the participant explicitly mentioned simplicity, or (2) the justification emphasized that the *single* disease accounted for *both* symptoms, thus suggesting that it was unnecessary to invoke two diseases when one would do the trick. Justifications were categorized as 'probability' if the participant claimed their choice was more probable or seemed more likely to be true. Both participants who computed the joint probability of D_2&D_3 and those who went on a subjective feeling of probability were included in this category. Finally, participants whose justifications could not be classified as simplicity or probability were included in the 'other' category. Often these justifications included a restatement of the question ("it seemed best" or "it seemed most satisfying") or an appeal to general intuition ("I went with my gut feeling").

As before, the disease names were counterbalanced, and the explanation choices were presented in random order. In addition, we counterbalanced the order of the presentation of the diseases such that half the participants read about D_1 first and half read about D_1 last. We used three different sets of symptoms.

Results and Conclusions Figure 1 indicates the percentage of participants choosing the simpler explanation in each frequency condition. Nearly all participants chose the simpler explanation when the probability ratio of D_1 to D_2&D_3 was close, but this number steadily declined as it became increasingly probable that an alien had D_2&D_3. Even when it was ten times more likely for the alien to have D_2&D_3, however, over a third of participants were still choosing the simpler explanation. Nor was this preference for the simpler explanation due to participants' inability to compute joint probabilities. The correlation between explanation choice and answering the math problem correctly was small and not significantly different from zero ($r = .12$, $p > .15$).
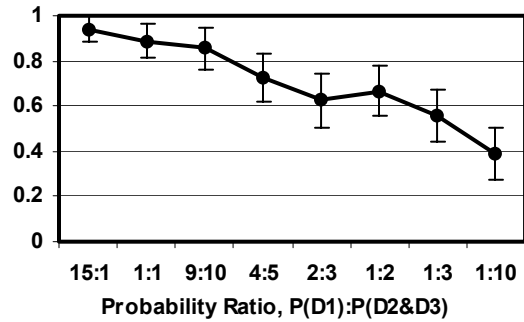

Figure 1: Percent of participants choosing simpler explanation as a function of the probability condition.

To better understand the data we conducted a logistic regression analysis. We used the natural log of the probability ratio as the predictor for the percentage of participants choosing the simpler explanation, as this choice results in a straightforward interpretation of the regression parameters. To understand why, it helps to consider how these parameters relate to the computations that would be performed by an idealized Bayesian agent. In our task, the ideal agent's data would result in a slope parameter of 1 and a constant of 0. A non-ideal agent could have a bias in favor of simplicity at either of two stages in the inference process, each corresponding to a parameter of the logistic function. A slope significantly less than 1 would suggest that the agent underweights the importance of probability: as evidence in favor of D_2&D_3 accumulates, the agent fails to reduce the probability of choosing D_1 accordingly. In contrast, a constant significantly different from zero reflects a bias at the level of the prior probability. The non-ideal agent could overweight, underweight, or appropriately weight probability information, but starts out with disproportionate confidence that D_1 is true.

The *probabilistic metric* and *trade-off hypotheses* make different predictions about the parameters of the logistic function resulting from this analysis. Specifically, the *probabilistic metric* hypothesis requires that the slope parameter be 1. If the preference for simplicity is represented in terms of probability, then probability information should be weighted appropriately. The constant, however, could be significantly different from 0. In contrast, the *trade-off* hypothesis makes no predictions about these parameters. Because simplicity and probability are

evaluated on different metrics, the bias could be reflected in either or both parameters.

The regression analysis resulted in a constant significantly different from zero, but a slope not significantly different from one. This provides some support for the *probabilistic metric* hypothesis. Specifically, the data suggest that as a group, participants think the simpler explanation is more likely than the complex alternative by a factor of about 4 (1.4 to 9, .95 confidence interval), and this belief influences what would be the prior probability in a Bayesian computation. When the probability ratio is 1:2, the percentage of subjects choosing the simpler explanation corresponds to the ideal Bayesian's posterior probability for D_1 at a frequency of 1:(2/4), and so on for the other values. As a result, participants require disproportionate evidence in favor of the complex explanation before it can rival the simpler alternative. Nonetheless, the slope of the regression suggests that participants incorporate probability information appropriately in making a decision.

We can also examine participants' beliefs about simplicity by looking at how they justified their explanation choices. When the simpler explanation was also more probable, a majority of participants justified choosing the simpler explanation by appeal to probability. However, 'simplicity' and 'other' explanations became increasingly common as the simpler explanation became less probable. These trends are illustrated in Figure 2, which indicates the percent of each justification type for the simpler explanation. Because there were few participants in some categories, the figure combines data from pairs of probability ratios.
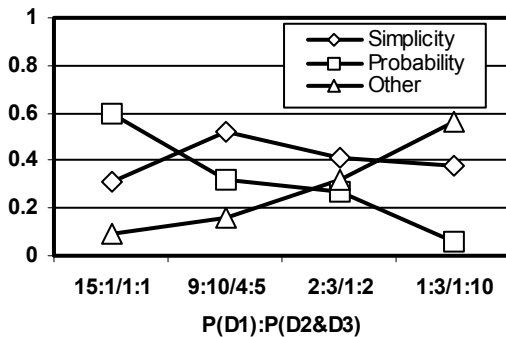

Figure 2: Distribution of justifications for choosing the simpler explanation.

Overall, the data suggest that many participants thought simpler explanations were more likely to be true. This is most apparent from the logistic regression analysis, but is also supported by the patterns of explanation justifications. In particular, many participants justified the choice of a simpler explanation by appeal to probability in conditions where the more complex explanation was as much as two times more likely. The data from the math problem suggest that this preference does not result from an inability to compute joint probabilities, but Experiment 3 considers and eliminates another alternative explanation.

Experiment 3: Independence Assumptions

In Experiment 2 we found that participants chose the simpler explanation well beyond the point at which a more complex alternative was more probable. However, the probability values against which we compared participants' choices were calculated on the assumption that diseases D_2 and D_3 are probabilistically independent—that is, that $P(D_2|D_3) = P(D_2)$ and $P(D_3|D_2) = P(D_3)$. As participants were told nothing about the dependence of the diseases, it's possible that they made a different assumption. In particular, if $P(D_2|D_3)$ is much smaller than $P(D_2)$, participants would be warranted in choosing D_1 on probabilistic grounds.

In Experiment 3 we were interested in determining participants' beliefs about the dependence of diseases. We also wanted to assess whether such beliefs influence explanatory preferences. To do so we found a domain involving dependence assumptions distinct from those for diseases, and examined whether more participants chose the complex explanation for this domain.

Methods Sixty-eight Boston-area undergraduate and summer school students participated. Twenty were in the *assumptions* condition, where we explicitly asked participants to provide a judgment of probabilistic dependence as follows:

Suppose there are two diseases with similar symptoms, D_1 and D_2. Do you think someone who has D_1 is more or less likely to have D_2 than someone who does not have D_1?
    Circle one: **More**   **Less**

In addition to asking about the dependence of diseases, we also wanted to find items with a different dependence assumption. We thus queried participants about books:

Suppose there are two books on similar topics, B_1 and B_2. Do you think someone who has read B_1 is more or less likely to have read B_2 than someone who has not read B_1?
    Circle one: **More**   **Less**

Participants in the *assumptions* condition saw both the disease and book questions, with the order counterbalanced.

The remaining 48 participants performed a task like Experiment 2 at the 2:3 probability ratio. However, half were asked about diseases, while the remaining half saw a formally identical question about books. Instead of reasoning about diseases causing symptoms, they were asked about books 'causing' knowledge of facts. For example, a passage read: "***The Zorgian Guide to Interplanetary Living*** contains the fact that **Planet Earth has an atmosphere** and the fact that **humans have two legs**. You know that about 50 aliens on Zorg have read ***The Zorgian Guide to Interplanetary Living***."

Results and Conclusions We first analyzed the data from the *assumptions* condition. Most participants (80%) claimed that having a disease makes someone *less* likely to have a similar disease, but 80% thought that having read a book makes someone *more* likely to have read a similar book.

These values were significantly different from chance, as well as being significantly different from each other ($\chi^2(1) = 14.4$, $p < .01$). Having thus established that people have different dependence assumptions about diseases and books, we went on to look at whether these assumptions affect explanatory preferences.

Replicating Experiment 2, we found that about half (46%) of participants chose the simpler explanation at the 2:3 probability ratio in the disease condition. In the book condition the results were identical, with 11 of 24 participants (46%) choosing the simpler explanation. There were also no differences between conditions in how participants justified their choice. The absence of a difference between the disease and book conditions suggests that beliefs about probabilistic dependence do not account for participants' preference for simpler explanations.

Experiment 4: Computer Replication

In the previous experiments participants were informed of the prevalence of each disease by being presented with a frequency. This method has two limitations. First, in the real world most frequency information is acquired through experience rather than a summary value, making the ecological validity of the task questionable. Second, having actual numbers allowed some participants to compute the joint probability of the diseases rather than relying on subjective judgments. For these reasons we decided to replicate the basic task in a computer format. Doing so also allowed us to examine whether explanatory preferences have consequences for perceived frequencies.

Methods One-hundred-and-eight Boston-area summer school and undergraduate students participated. The task was like Experiment 2, but on the computer. Instead of being told the prevalence of the diseases, for each disease participants saw ten screens containing a total of 75 aliens, some of which were marked as having a particular disease. In this way equivalent frequency information was communicated.

Participants were in one of four frequency conditions corresponding to probability ratios of 15:1, 9:10, 1:2 and 1:10, with 27 participants per condition. After being presented with the cover story and frequency information, participants were asked to choose the most satisfying explanation for the alien's symptoms and, as before, selected an answer among six options, which included every disease alone and each pairwise combination. They were then asked to explain their choice and to estimate the frequency of each disease in the Zorg population.

Counterbalancing and randomization was as in Experiment 2, with the additional control that the order of presentation of the disease frequencies was varied according to a Latin square.

Results and Conclusions The overall explanatory preferences in the computer task replicated those of Experiment 2, suggesting that the questionnaire format was methodologically sound (see Figure 3). Virtually all participants chose the simpler explanation when it was more likely, but nearly half continued to prefer the simpler explanation when it was as much as ten times more likely that the alien had two diseases.
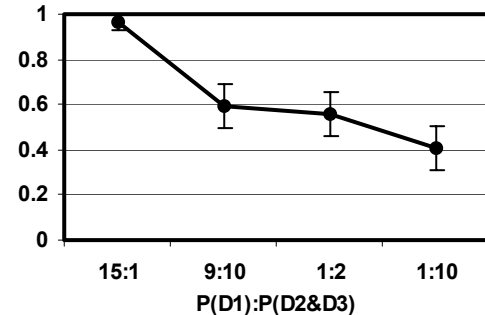

Figure 3: Percent of subjects choosing simpler explanation in computer task.

We also analyzed justifications for choosing the simpler explanation, using the coding scheme from Experiment 2. Not surprisingly, as the simpler explanation became less probable, a larger proportion of participants invoked simplicity rather than probability in their justifications (see Figure 4).
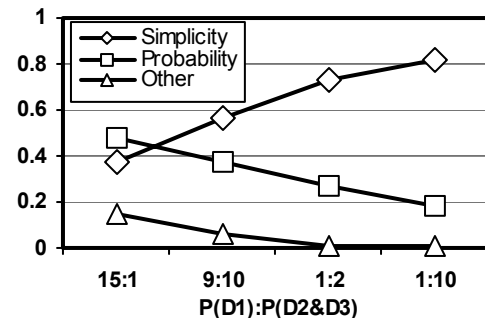

Figure 4: Distribution of justifications for choosing the simpler explanation in the computer task.

Using a computer task allowed us to examine an aspect of simplicity we couldn't address in the questionnaire format, namely how explanatory choices affect perceived frequencies. Figure 5 presents participants' estimates of the percentage of the Zorg population with each of D_1, D_2, and D_3. The average estimates are shown as a function of both frequency condition and explanation choice, with participants who chose the simpler, one-cause explanation distinguished from those who chose the more complex, two-cause explanation. Solid lines indicate the actual percentage of aliens with each disease.

While subjects were remarkably accurate overall, the data for D_1 suggest that those participants who chose the simple, one-cause explanation when it was less probable systematically overestimated the frequency of D_1. In both the 1:2 and 1:10 frequency conditions, the average estimate of participants who chose the simpler explanation were
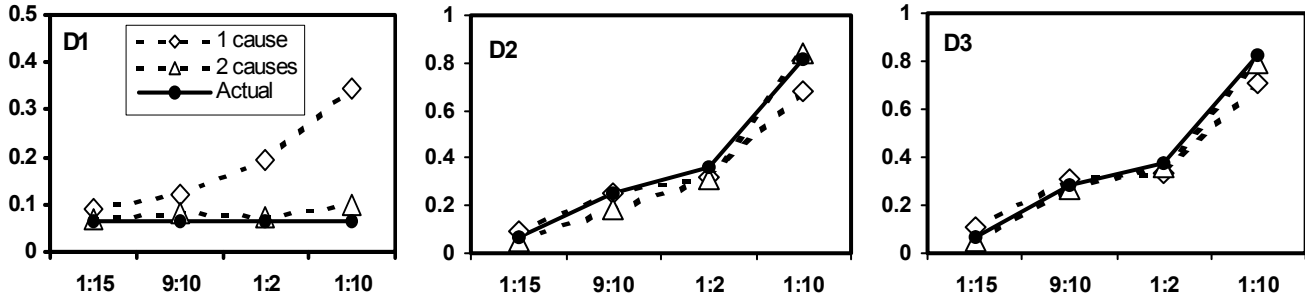
Figure 5: Average frequency estimates for each disease as a function of probability ratio and explanation choice.

significantly higher than that of participants who chose the more complex alternative ($p < .05$). One possibility, however, is that some subjects confused the frequency of D_1 with either D_2 or D_3, which would result in inflated D_1 estimates. If this were true we would expect to see systematic underestimation of D_2 and D_3. Moreover, only one subject provided a higher estimate for D_1 than either D_2 or D_3. This suggests that the overestimation of D_1 is not due to mismatching the frequencies and their corresponding diseases.

Another explanation for the D_1 overestimation is that participants who chose the simpler explanation were bad at estimating frequencies, and for this reason based their explanation choice on simplicity. This possibility is ruled out by the frequency estimation data for D_2 and D_3, where there were no differences between the estimates of participants who chose one or two cause explanations.

These data suggest that participants who chose the simpler explanation systematically overestimated the frequency of D_1 *as a result* of their explanation choice. However, it could be that some participants overestimated D_1, which in turn lead them both to choose the simpler explanation and to indicate a high prevalence of D_1. Evidence that the former interpretation is the correct one comes from the fact that participants never systematically overestimate D_1 in the 1:15 condition, when simplicity and probability converged on the same explanation.

Conclusions

We began by considering whether people prefer simpler explanations, and whether this preference is supported by a belief that simpler explanations are more likely to be true. We found overwhelming evidence for the claim that people do prefer simpler explanations, at least where simplicity is understood in terms of number of causes. Participants consistently chose a simpler explanation when provided no information about probability, and required a disproportionate amount of probability information in order to override this preference.

These findings are consistent with the idea that people believe simpler explanations are more likely to be true, albeit implicitly. Many subjects explicitly justified their choice of a simpler explanation by appeal to probability, but more telling is the fact that participants evaluated simplicity

and probability as if they were commensurable quantities. The results from Experiment 2 tentatively support the *probabilistic metric* hypothesis over the *trade-off* hypothesis: people do prefer simpler explanations, but this bias manifests as a reweighing of priors rather than a failure to appropriately incorporate probability information.

The intimate relationship between simplicity and probability is most dramatically illustrated by the finding that committing to an improbable, simple explanation results in the systematic distortion of perceived frequencies. This result indicates that explanatory choices can have consequences for probabilistic judgments, and suggests that the study of explanation can provide a unique window into the mechanisms by which beliefs about the world influence decisions.

References

Chater, N. & Vitanyi, P. (2003). Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Science, 7(1)*, 19-22.

Jeffreys, W. H. & Berger, J.O. (1992). Ockham's razor and Bayesian analysis. *American Scientist*, *80*, 64-72.

Lagnado, D. (1994). *The Psychology of Explanation: A Bayesian Approach*. Masters Thesis. Schools of Psychology and Computer Science, University of Birmingham.

Newton I. (1953/1686). *Philosophiae Naturalis Principia Mathematica*. Reprinted in H. Thayer (Ed.) *Newton's Philosophy of Nature*. New York: Hafner.

Sober, E. (Forthcoming). Parsimony. In S. Sarkar (Ed.), *The Philosophy of Science—an Encyclopedia.* New York: Routledge.