How to Never be Wrong

Samuel J. Gershman
Department of Psychology and Center for Brain Science
Harvard University

Human beliefs have remarkable robustness in the face of disconfirmation. This robustness is often explained as the product of heuristics or motivated reasoning. However, robustness can also arise from purely rational principles when the reasoner has recourse to *ad hoc* auxiliary hypotheses. Auxiliary hypotheses primarily function as the linking assumptions connecting different beliefs to one another and to observational data, but they can also function as a "protective belt" that explains away disconfirmation by absorbing some of the blame. The present article traces the role of auxiliary hypotheses from philosophy of science to Bayesian models of cognition and a host of behavioral phenomena, demonstrating their wide-ranging implications.

> *"No theory ever agrees with all the facts in its domain, yet it is not always the theory that is to blame. Facts are constituted by older ideologies, and a clash between facts and theories may be proof of progress."* (Feyerabend, 1975)

Introduction

Since the discovery of Uranus in 1781, astronomers were troubled by certain irregularities in its orbit, which appeared to contradict the prevailing Newtonian theory of gravitation. Then, in 1845, Le Verrier and Adams independently completed calculations showing that these irregularities could be entirely explained by the gravity of a previously unobserved planetary body. This hypothesis was confirmed a year later through telescopic observation, and thus an 8th planet (Neptune) was added to the solar system. Le Verrier and Adams succeeded on two fronts: they discovered a new planet, and they rescued the Newtonian theory from disconfirmation.

Neptune is a classic example of what philosophers of science call an *ad hoc auxiliary hypothesis* (Hempel, 1966; Popper, 1959). All scientific theories make use of auxiliary assumptions that allow them to interpret experimental data. For example, an astronomer makes use of optical assumptions to interpret telescope data, but one would not say that these assumptions are a core part of an astronomical theory; they can be replaced by other assumptions as the need arises (e.g., when using a different measurement device), without threatening the integrity of the theory. An auxiliary assumption becomes an *ad hoc* hypothesis when it entails unconfirmed claims that are specifically designed to accommodate disconfirmatory evidence.

Ad hoc auxiliary hypotheses have long worried philosophers of science, because they suggest a slippery slope towards unfalsifiability (Harding, 1976). If any theory can be rescued in the face of disconfirmation by changing auxiliary assumptions, how can we tell good theories from bad theories? While Le Verrier and Adams were celebrated for their discovery, many other scientists were less fortunate. For example, in the late 19th century, Michelson and Morley reported experiments apparently contradicting the prevailing theory that electromagnetic radiation is propagated through a space-pervading medium (ether). FitzGerald and Lorentz attempted to rescue this theory by hypothesizing electrical effects of ether that were of exactly the right magnitude to produce the Michelson and Morley results. Ultimately, the ether theory was abandoned, and Popper (1959) derided the FitzGerald-Lorentz explanation as "unsatisfactory" because it "merely served to restore agreement between theory and experiment."

Ironically, Le Verrier himself was misled by an *ad hoc* auxiliary hypothesis. The same methodology that had served him so well in the discovery of Neptune failed catastrophically in his "discovery" of Vulcan, a hypothetical planet postulated to explain excess precession in Mercury's orbit. Le Verrier died convinced that Vulcan existed, and many astronomers subsequently reported sightings of the planet, but the hypothesis was eventually discredited by Einstein's theory of general relativity, which accounted precisely for the excess precession without recourse to an additional planet.

The basic problem posed by these examples is how to assign credit or blame to central hypotheses vs. aux-

Address for correspondence:
52 Oxford St., Room 295.05
Cambridge, MA 02138
e-mail: gershman@fas.harvard.edu

iliary hypotheses. An influential view, known as the Duhem-Quine thesis (reviewed in the next section), asserts that this credit assignment problem is insoluble—central and auxiliary hypotheses must face observational data "as a corporate body" (Quine, 1951). This thesis implies that theories will be resistant to disconfirmation as long as they have recourse to *ad hoc* auxiliary hypotheses.

Psychologists recognize such resistance as a ubiquitous cognitive phenomenon, commonly viewed as one among many flaws in human reasoning (Gilovich, 1991). However, as the Neptune example attests, such hypotheses can also be instruments for discovery. The purpose of this paper is to discuss how a Bayesian framework for induction deals with *ad hoc* auxiliary hypotheses (Dorling, 1979; Earman, 1992; Howson and Urbach, 2006; Strevens, 2001), and then to leverage this framework to understand a range of phenomena in human cognition. According to the Bayesian framework, resistance to disconfirmation can arise from rational belief updating mechanisms, provided that an individual's "intuitive theory" satisfies certain properties: a strong prior belief in the central hypothesis, coupled with an inductive bias to posit auxiliary hypotheses that place high probability on observed anomalies. The question then becomes whether human intuitive theories satisfy these properties, and several lines of evidence suggest the answer is yes.[1] In this light, humans are surprisingly rational. Human beliefs are guided by strong inductive biases about the world. These biases enable the development of robust intuitive theories, but can sometimes lead to preposterous beliefs.

Underdetermination of theories: the Duhem-Quine thesis

Theories (both scientific and intuitive) are webs of interconnected hypotheses about the world. Thus, one often cannot confirm or disconfirm one hypothesis without affecting the validity of the other hypotheses. How, then, can we establish the validity of an individual hypothesis? Duhem (1954) brought this issue to the foreground in his famous treatment of theoretical physics:

> The physicist can never subject an isolated hypothesis to experimental test, but only a whole group of hypotheses; when the experiment is in disagreement with his predictions, what he learns is that at least one of the hypotheses constituting this group is unacceptable and ought to be modified; but the experiment does not designate which one should be changed. (p. 187)

While Duhem restricted his attention to theoretical physics, Quine (1951) took the same point to its logical extreme, asserting that *all* beliefs about the world are underdetermined by observational data:

> The totality of our so-called knowledge or beliefs, from the most casual matters of geography and history to the profoundest laws of atomic physics or even of pure mathematics and logic, is a man-made fabric which impinges on experience only along the edges. Or, to change the figure, total science is like a field of force whose boundary conditions are experience. A conflict with experience at the periphery occasions readjustments in the interior of the field. But the total field is so underdetermined by its boundary conditions, experience, that there is much latitude of choice as to what statements to reevaluate in the light of any single contrary experience. No particular experiences are linked with any particular statements in the interior of the field, except indirectly through considerations of equilibrium affecting the field as a whole. (p. 42-43)

In other words, one cannot unequivocally identify particular beliefs to revise in light of surprising observations. Quine's conclusion was stark: "The unit of empirical significance is the whole of science" (p. 42).

Some philosophers have taken undetermination to invite a radical critique of theory-testing. if evidence cannot adjudicate between theories, then non-empirical forces, emanating from the social and cultural environment of scientists, must drive theory change. For example, the "research programmes" of Lakatos (1976) and the "paradigms" of Kuhn (1962) were conceived as explanations of why scientists often stick to a theory despite disconfirming evidence, sometimes for centuries. Lakatos posited that scientific theories contain a hard core of central theses that are immunized from refutation by a "protective belt" of auxiliary hypotheses. On this view, science does not progress by falsification of individual theories, but rather by developing a *sequence* of theories that progressively add novel predictions, some of which are corroborated by empirical data.

While the radical consequences of underdetermination have been disputed (e.g., Grünbaum, 1962; Laudan, 1990), the problem of credit assignment remains

[1] As a caveat, we should keep in mind that whether a particular intuitive theory satisfies these properties will naturally vary across domains and an individual's experience.

a fundamental challenge for the scientific enterprise. I now turn to a Bayesian approach to induction that attempts to answer this challenge.

The Bayesian answer to underdetermination

Probability theory offers a coherent approach to credit assignment (Howson and Urbach, 2006). Instead of assigning all credit to either central or auxiliary hypotheses, probability theory dictates that credit should be apportioned in a graded manner according to the "responsibility" each hypothesis takes for the data. More formally, let h denote the central hypothesis, a denote the auxiliary hypothesis, and d denote the data. After observing d, the prior probability of the conjunct ha, $P(ha)$, is updated to the posterior distribution $P(ha|d)$ according to Bayes' rule:

$$P(ha|d) = \frac{P(d|ha)P(ha)}{P(d|ha)P(ha) + P(d|\neg(ha))P(\neg(ha))}, \quad (1)$$

where $P(d|ha)$ is the likelihood of the data under ha, and $\neg(ha)$ denotes the negation of ha.

The sum rule of probability allows us to ascertain the updated belief about the central hypothesis, marginalizing over all possible auxiliaries:

$$P(h|d) = P(ha|d) + P(h\neg a|d). \quad (2)$$

Likewise, the marginal posterior over the auxiliary is given by:

$$P(a|d) = P(ha|d) + P(\neg ha|d). \quad (3)$$

This formulation is the crux of the Bayesian answer to underdetermination (Dorling, 1979; Earman, 1992; Howson and Urbach, 2006; Strevens, 2001). A Bayesian scientist does not wholly credit either the central or auxiliary hypotheses, but rather distributes the credit according to the marginal posterior probabilities.

This analysis does not make a principled distinction between central and auxiliary hypotheses: they act conjunctively, and are acted upon in the same way by the probability calculus. What ultimately matters for distinguishing them, as illustrated below, is the relative balance of evidence for the different hypotheses. Central hypotheses will typically be more entrenched due to a stronger evidential foundation, and thus auxiliary hypotheses will tend to be the element's of Quine's "total field" that readjust in the face of disconfirmation.

I will not address here the philosophical controversies that have surrounded the Bayesian analysis of auxiliary hypotheses (Fitelson and Waterman, 2005; Mayo, 1997). My goal is not to establish the normative adequacy of the Bayesian analysis, but rather to explore its implications for cognition—in particular, how it helps us understand resistance to belief updating.

Following Strevens (2001), I illustrate the dynamics of belief by assuming that the data d has its impact on the posterior probability of the central hypothesis h solely through its falsification of the conjunct ha[2]:

$$P(h|d) = P(h|\neg(ha)). \quad (4)$$

In other words, the likelihood is 0 for ha and 1 for all other conjuncts. Under this assumption, Strevens obtains the following expression:

$$P(h|d) = \frac{1 - P(a|h)}{1 - P(a|h)P(h)}P(h). \quad (5)$$

This expression has several intuitive properties, illustrated in Figure 1. As one would expect, the posterior probability of h always decreases following disconfirmatory data d. The decrease in the posterior probability is inversely proportional to $P(h)$ and directly proportional to $P(a|h)$.[3] Thus, a central hypothesis with high prior probability relative to the auxiliary hypothesis [i.e., high $P(h)/P(a|h)$] will be relatively robust to disconfirmation, pushing blame onto the auxiliary. But if the auxiliary has sufficiently high prior probability, the central hypothesis will be forced to absorb the blame. It is important to see that the robustness to disconfirmation conferred by a strong prior is not a bias due to motivated reasoning (Kunda, 1990)—it is a direct consequence of rational inference. This will be a key point reiterated throughout the paper.[4]

One might wonder how this analysis determines whether an auxiliary hypothesis is *ad hoc* or not. The answer is that it doesn't: the only distinguishing features of hypotheses are their prior probabilities and their likelihoods. Thus, on this account "*ad hoc*" is simply a descriptive label that we use to individuate hypotheses that have low prior probability and high likelihoods. By the same token, a "good" versus "bad" *ad hoc* auxiliary hypothesis is determined entirely by the prior and likelihood.

[2]Strevens (2001) notes that this expression does not hold if the data affect the posterior in ways other than falsifying the conjunct ha, although such scenarios are probably rare.

[3]We can think of this conditional prior as specifying strength of belief in an auxiliary given that one already believes a particular central hypothesis. In other words, it assumes that different central hypotheses invoke different distributions over auxiliaries. This seems intuitive insofar as auxiliaries will tend to be highly theory-specific (you don't hypothesize auxiliaries about baseball when contemplating cosmology).

[4]This is not to deny that some forms of motivated reasoning exist, but only to assert particular ways in which robustness to disconfirmation arises from rational inference.

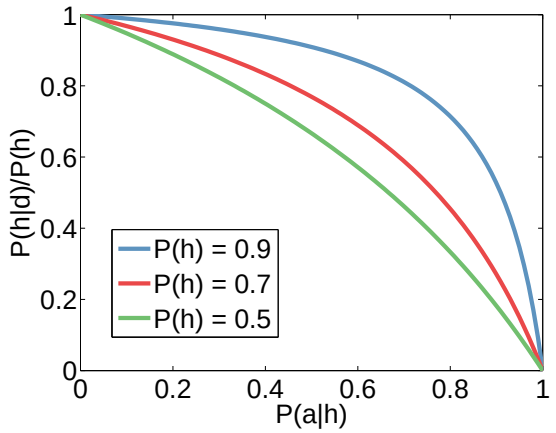

Figure 1. **Simulations**. Ratio of posterior to prior probability of the central hypothesis h as a function of the probability of the auxiliary hypothesis a given h, plotted for three different priors for the central hypothesis. Adapted from Strevens (2001).

Robustness of intuitive theories

One strong assumption underlying this analysis is worth highlighting, namely that the likelihood of $h\neg a$, marginalizing over all alternative auxiliaries (a_k), is equal to 1:

$$P(d|h\neg a) = \sum_{a_k \neq a} P(d|ha_k)P(a_k) = 1. \quad (6)$$

I will refer to this as the *consistency assumption*, because it states that only auxiliary hypotheses that are highly consistent with the data will have non-zero probability. Mathematically, this means that $P(a_k) > 0$ if and only if $P(d|ha_k) = 1$. *Ad hoc* auxiliary hypotheses, by design, have the property that $P(d|ha_k) \approx 1$. But why should these hypotheses be preferred over others? One way to justify this assumption is to stipulate that there is uncertainty about the parameters of the distribution over auxiliary hypotheses. The prior over these parameters can express a preference for redistributing probability mass (i.e., assigning credit) in particular ways once data are observed.

Concretely, let θ denote the parameter vector of the multinomial distribution over auxiliaries. Because we have uncertainty about θ in addition to h and a, we need to marginalize over θ to obtain the posterior $P(h|d)$:

$$P(h|d) = \int_\theta P(h|d,\theta)P(\theta|d)d\theta. \quad (7)$$

As detailed in the Appendix, choosing $P(\theta)$ to be a sparse Dirichlet distribution has the effect that $P(d|h\neg a) \approx 1$. A sparse Dirichlet distribution places most of its probability mass on multinomial distributions with low entropy (i.e., those that favor a small set of auxiliary hypotheses). After observing d, the marginal distribution $P(a_k|d)$ will place most of its probability mass on auxiliary hypotheses that are consistent with the data. In other words, the assumption of sparsity licenses us to discount all the auxiliary hypotheses that are inconsistent with the data. The remaining auxiliaries may appear as though they are *ad hoc*, but in fact they are the only ones that survive the cull of rational inference.

In addition to sparsity, the consistency assumption requires *deterministic* hypotheses: $P(d|ha_k)$ must be close to 1 if a_k is to be considered plausible (see Appendix). If hypotheses are allowed to probabilistically predict the data, then $P(d|h\neg a) < 1$. In summary, sparsity and determinism jointly facilitate the robustness of theories. In this section, I will argue that these properties characterize human intuitive theories.

Sparsity

The sparsity assumption—that only a few auxiliary hypotheses have high probability—has appeared throughout cognitive science in various guises. Klayman and Ha (1987) posited a *minority phenomenon* assumption, according to which the properties that are characteristic of a hypothesis tend to be rare. For example, AIDS is rare in the population but highly correlated with HIV; hence observing that someone has AIDS is highly informative about whether they have HIV. Klayman and Ha (1987) invoked this assumption to justify the "positive test strategy" prevalent in human hypothesis testing. If people seek confirmation for their hypotheses, then failure to observe the confirmatory evidence will provide strong evidence against the hypothesis under the minority phenomenon assumption. Oaksford and Chater (1994) used the same idea (what they called the *rarity assumption*) to explain the use of the positive test strategy in the Wason card selection task. Violations of the sparsity assumption, or contextual information that changes perceived sparsity, causes people to shift away from the positive test strategy (Hendrickson et al., 2016; McKenzie et al., 2001). Experiments on hypothesis evaluation tell a similar story: the evidential impact of observations is greater when they are rare (Mckenzie and Mikkelsen, 2000; McKenzie and Mikkelsen, 2007), consistent with the assumption that hypotheses are sparse.

Beyond hypothesis testing and evaluation, evidence suggests that people tend to generate sparse hypotheses when presented with data. For example, Perfors and Navarro (2009) asked participants to generate hy-

pothetical number concepts applicable to the range $[1, 1000]$, and found that most of these hypotheses were sparse. For example, a common hypothesis was prime numbers, with a sparsity of 0.168 (i.e., 16.8% of numbers in $[0, 1000]$ are primes). Overall, 83% of the generated hypotheses had a sparsity level of 0.2 or less.[5]

Sparsity has also figured prominently in theories of perception. Olshausen and Field (1996) accounted for the tuning properties of receptive fields in primary visual cortex by assuming that they represent a sparse set of image components. Similar sparse coding ideas have been applied to auditory (Hromádka et al., 2008) and olfactory (Poo and Isaacson, 2009) cortical representations. Psychologists have likewise posited that humans parse complex objects into a small set of latent components with distinctive visual features (Austerweil and Griffiths, 2013; Biederman, 1987).

Is sparsity a reasonable assumption? Navarro and Perfors (2011) attempted to answer this question by demonstrating that (under some fairly generic assumptions) sparsity is a consequence of *family resemblance*: hypotheses tend to generate data that are more similar to one another than to data generated by other hypotheses. For example, members of the same natural category tend to have more overlapping features relative to members of other natural categories (Rosch, 1978). Navarro and Perfors (2011) further showed that natural categories are empirically sparse. Thus, the sparsity assumption may be inevitable if hypotheses describe natural categories.

Determinism

The determinism assumption—that hypotheses tend to generate data near-deterministically—is well-supported as a property of intuitive theories. Some of the most compelling evidence comes from studies of children showing that children will posit a latent cause to explain surprising events, rather than attribute the surprising event to inherent stochasticity (Buchanan and Sobel, 2011; Muentener and Schulz, 2014; Saxe et al., 2005; Schulz and Sommerville, 2006; Wu et al., 2015). For example, Schulz and Sommerville (2006) presented 4-year-olds with a stochastic generative cause and found that the children inferred an inhibitory cause to "explain away" the stochasticity. Children also expect latent agents to be the cause of surprising motion events, even in the absence of direct evidence for an agent (Saxe et al., 2005). Like children, adults also appear to prefer deterministic hypotheses (Frosch and Johnson-Laird, 2011; Mayrhofer and Waldmann, 2015).[6] The prevalent use of the positive test strategy in information selection has also been justified using the determinism assumption (Austerweil and Griffiths, 2011).

Lu et al. (2008) have proposed a "generic prior" for causal strength that combines the sparsity and determinism principles. *A priori*, causes are expected to be few in number and potent in their generative or preventative effects. Lu et al. (2008) showed quantitatively that this prior, when employed in a Bayesian framework for causal induction, provides a good description of human causal inferences.[7] Buchanan et al. (2010) developed an alternative deterministic causal model based on an edge replacement process, which creates a branching structure of stochastic latent variables. This model can explain violations of conditional independence in human judgments in terms of the correlations induced by the latent variables.

In summary, sparsity and determinism appear to be central properties of intuitive theories. These properties offer support for the particular Bayesian analysis of auxiliary hypotheses elaborated above, according to which robustness of theories derives from the ability to explain away disconfirmatory data by invoking auxiliary hypotheses.

Implications

Having established the plausibility of the Bayesian analysis, we now explore some of its implications for human cognition. The central theme running through all of these examples is that the evidential impact of observations is contingent on the auxiliary hypotheses one holds; changing one's beliefs about auxiliary hypotheses will change the interpretation of observations. Thus, observations that appear to contradict a central hypothesis can be "explained away" by changing auxiliary hypotheses, and this change is licensed by the Bayesian analysis under the specific circumstances detailed above. If, as I have argued, intuitive theories have the right sort of properties to support this "protective belt" of auxiliary hypotheses (cf. Lakatos, 1976), then we should expect robustness to disconfirmation across many domains.

---

[5] Note that there are a number of reasons why people might generate sparse hypotheses besides having a sparse prior, such as computational limits (cf. Dasgupta et al., 2017).

[6] Some evidence suggests that people can adaptively determine which causal theory (deterministic or probabilistic) is most suitable for a given domain (Griffiths and Tenenbaum, 2009).

[7] Yeung and Griffiths (2015) presented empirical evidence favoring a preference for (near) determinism but not sparsity, though other experiments have suggested that both sparsity and determinism are required to explain human causal inferences (Powell et al., 2016).

Before proceeding, it is important to note that many of the phenomenon surveyed below can also be explained by other theoretical frameworks, such as motivated cognition (Kunda, 1990). The purpose of this section is not to develop a watertight case for the Bayesian framework—which would require more specific model specifications for different domains and new experiments to test rival predictions—but rather to show that evidence for robustness to disconfirmation does not by itself indicate irrationality; it is possible to conceive of a perfectly rational agent who exhibits such behavior. Whether humans really are rational in this way is an unresolved empirical question.[8]

The theory-ladenness of observation

> *"It is quite wrong to try founding a theory on observable magnitudes alone. In reality the very opposite happens. It is the theory which decides what we can observe." (Albert Einstein)*

Drawing a comparison between the history of science and perceptual psychology, Kuhn (1962) argued that observation reports are not theory-neutral: "What a man sees depends both upon what he looks at and also upon what his previous visual-conceptual experience has taught him to see" (p 113). For example, subjects who put on goggles with inverting lenses see the world upside-down, but after a period of profound disorientation lasting several days, their perception adapts and they see the world right-side-up (Stratton, 1897). Thus, the very same retinal image produces starkly different percepts depending on the preceding perceptual history.

More important for Kuhn's argument are examples where percepts, or at least their semantic interpretations, are influenced by the observer's conceptual framework:

> Looking at a contour map, the student sees lines on paper, the cartographer a picture of a terrain. Looking at a bubble-chamber photograph, the student sees confused and broken lines, the physicist a record of familiar subnuclear events. Only after a number of such transformations of vision does the student become an inhabitant of the scientist's world, seeing what the scientist sees and responding as the scientist does. The world that the student then enters is not, however, fixed once and for all by the nature of the environment, on the one hand, and of science, on the other. Rather, it is determined jointly by the environment and the particular normal-scientific tradition that the student has been trained to pursue. (Kuhn, 1962, pp. 111–112)

This is essentially a restatement of the view, going back to Helmholtz (1867), that perception is a form of "unconscious inference" or "problem-solving" (Gregory, 1970; Rock, 1983) and formalized by modern Bayesian theories of perception (Knill and Richards, 1996).[9]

There is one particular form of theory-ladenness that will concern us here, where changes in auxiliary hypotheses alter the interpretation of observations. Disconfirmation can be transformed into confirmation (e.g., the example of Neptune), or vice versa. When Galileo first reported his observations of mountains on the moon, the critical response focused not on the observations *per se* but on the auxiliary assumptions mediating their validity. Since the telescope was an unfamiliar measurement device, the optical theory underlying its operation was not taken for granted. In fact, it was non-trivial even to verify Galileo's observations, because many of the other telescopes available in 1610 were of insufficient quality to resolve the same lunar details observed by Galileo. Thus, it was possible at that time to dispute the evidential impact of Galileo's observations for astronomical theories (see Bovens and Hartmann, 2002, for a detailed analysis of how beliefs about the unreliability of measurement instruments affects reasoning about auxiliary hypotheses).

Although Galileo's observations were ultimately vindicated, there are other historical examples in which observations were ultimately discredited. For example, Rutherford and Pettersson conducted similar experiments in the 1920s on the emission of charged particles under radioactive bombardment. Pettersson's assistants observed flashes on a scintillation screen (evidence for emission) whereas Rutherford's assistants did not. The controversy was subsequently resolved when Rutherford's colleague, James Chadwick, demonstrated that Pettersson's assistants were unreliable: they reported indistinguishable rates of flashes

[8] Indeed, there has been a vigorous debate in psychology about the validity of Bayesian rationality as a model of human cognition (e.g., Jones and Love, 2011). Here I am merely asking the reader to consider the conditional claim that *if* people are Bayesian with sparse and deterministic intuitive theories, then they would exhibit robustness to disconfirmation.

[9] It is important to distinguish this view from the stronger thesis that no theory-neutral stage of perceptual analysis exists (e.g., Churchland, 1979). As pointed out by Fodor (1984), we can accept that the semantic interpretation of percepts is theory-dependent without abandoning the possibility that there are *some* cognitively impenetrable aspects of perception.

even under experimental conditions where no particles could have been emitted. The strategy of debunking claims by undermining auxiliary hypotheses has been used effectively throughout scientific history, from Benjamin Franklin's challenge of Mesmer's "animal magnetism" to the revelation that observations of neutrinos exceeding the speed of light were due to faulty detectors.[10]

It is tempting to see a similar strategy at work in contemporary political and scientific debate. In response to negative news coverage, the Trump administration promulgated the idea that the mainstream media is publishing "fake news"—i.e., reports that are inaccurate, unreliable, or biased. This strategy is powerful because it does not focus on the veracity of any one report, but instead attempts to undermine faith in the entire "measurement device." A similar strategy was used for many years by creationists to undermine faith in evolutionary biology, by the tobacco industry to undermine faith in scientific studies of smoking's health effects, and by the fossil fuel industry to undermine faith in climate science. By "teaching the controversy," these groups attempt to dismantle the auxiliary hypotheses on which the validity of science relies. For example, the release of stolen e-mails from the Climatic Research Unit at the University of East Anglia suggested an alternative auxiliary—selective reporting or manipulation of data—that could explain away evidence for human-induced climate change. Indeed, a subsequent survey of Americans showed that over half agreed with the statements "Scientists changed their results to make global warming appear worse than it is" and "Scientists conspired to suppress global warming research they disagreed with" (Leiserowitz et al., 2013).

A well-studied form of theory-ladenness is the phenomenon of *belief polarization*: individuals presented with the same data will sometimes update their beliefs in opposite directions. In a classic experiment, Lord et al. (1979) asked supporters and opponents of the death penalty to read about two fictional studies—one supporting the effectiveness of the death penalty as a crime deterrent, and one supporting its ineffectiveness. Subjects who supported the death penalty subsequently strengthened their belief in the effectiveness of the death penalty after reading the two studies, whereas subjects who opposed the death penalty subsequently strengthened their belief in its ineffectiveness. A large body of empirical work on belief polarization was interpreted by many social psychologists as evidence of irrational belief updating (e.g., Kunda, 1990; Nisbett and Ross, 1980). However, another possibility is that belief polarization might arise from different auxiliary hypotheses about the data-generating process (Cook and Lewandowsky, 2016; Jaynes, 2003; Jern et al., 2014; Koehler, 1993). For example, Jern et al. (2014) showed how the findings of Lord et al. (1979) could be accounted for within a rational Bayesian framework. If participants assume the existence of research bias (distortion or selective reporting of findings to support a preconceived conclusion), then reading a study about the ineffectiveness of the death penalty may strengthen their belief in research bias, correspondingly increasing their belief in the effectiveness of the death penalty. Similarly, Cook and Lewandowsky (2016) demonstrated that beliefs in bias of scientific reporting can lead to discounting of climate change evidence. One lesson to draw from these examples is that effective persuasion requires more than simply conveying information confirming or disconfirming central hypotheses; it requires alteration of the auxiliary hypotheses that refract information, rendering perception theory-laden.

Optimism and controllability

Many individuals exhibit a systematic "optimism bias" (Sharot, 2011), overestimating the likelihood of positive events in the future.[11] This bias affects beliefs about many real-world domains, such as the probability of getting divorced or being in a car accident. One of the puzzles of optimism is how it can be maintained; even if we start with initial optimism (cf. Stankevicius et al., 2014), why doesn't reality force our beliefs to eventually calibrate themselves?

A clue to this puzzle comes from evidence that people tend to update their beliefs more in response to positive feedback compared to negative feedback (Eil and Rao, 2011; Sharot and Garrett, 2016). Eil and Rao (2011) dubbed this the "good news-bad news effect." For example, Eil and Rao asked subjects to judge the rank of their IQ and physical attractiveness and then received feedback (a pairwise comparison with a randomly selected subject in the same experiment). While subjects conformed to Bayesian updating when they received positive feedback (i.e., when their rank was

[10] How can this debunking strategy succeed when theorists can produce new auxiliary hypotheses *ad infinitum*? The Bayesian analysis makes provision for this: new auxiliaries will only be considered if they have appreciable probability, $P(a|h)$, relative to the prior, $P(h)$.

[11] The generality of this effect has been the subject of controversy, with some authors Shah et al. (2016) finding no evidence for an optimism bias. However, these null results have themselves been controversial: correcting confounds in the methodology (Garrett and Sharot, 2017), and using model-based estimation techniques (Kuzmanovic and Rigoux, 2017), have indicated a robust optimism bias.

better than the comparand), they systematically discounted the negative feedback. Similar results have been found using a variety of feedback types (Korn et al., 2012; Lefebvre et al., 2017; Sharot et al., 2011).

One reason people may discount negative feedback is that they wish to blunt its "sting" (Eil and Rao, 2011; Köszegi, 2006). Consistent with this account, Eil and Rao found that subjects who believed that their ranks were near the bottom of the distribution were willing to pay to avoid learning their true rank. An alternative account, drawing from our Bayesian analysis of auxiliary hypotheses, is that people are being fully Bayesian, but their internal model is different from the one presupposed by Eil and Rao. Specifically, let h denote the hypothesis that a person is "high rank," and let a denote the auxiliary hypothesis that the feedback is "valid" (i.e., from an unbiased source). It is intuitive that subjects might discount negative feedback by positing invalid evidence sources; for example, if a person judges you to be unattractive, you could discount this feedback by positing that this person is exceptionally harsh (judges everyone to be unattractive) or is having a bad day.

Suppose we have two people who have the same prior on validity, $P(a|h)$, but different priors on their rank, $P(h)$. The Bayesian analysis developed above (see Figure 1) predicts that the person who assigns higher prior probability to being high rank will update less in response to negative feedback. Consistent with this prediction, individuals with higher dispositional optimism were more likely to maintain positive expectations after experiencing losses in a gambling task (Gibson and Sanbonmatsu, 2004). The Bayesian analysis also predicts that two people with different priors on validity but the same priors on rank will exhibit different patterns of asymmetric updating, with the weaker prior on validity leading to greater discounting of negative feedback. In support of this prediction, Gilovich and colleagues (Gilovich, 1983; Gilovich and Douglas, 1986) found that people who observed an outcome that appeared to have arisen from a statistical "fluke" were more likely to discount this outcome when it was negative, presumably since the feedback was perceived to be invalid. The same kind of discounting can lead to overconfidence in financial markets, where investors are learning about their abilities; by taking too much credit for their gains and not enough for their losses, they become overconfident (Gervais and Odean, 2001).

A related phenomenon arises in studies of cheating and lying (see Gino et al., 2016, for a review). When people obtain a favorable outcome through unscrupulous means, they tend to attribute this success to their personal ability. For example, Chance et al. (2011) administered an IQ test to participants that included an answer key at the bottom so that they could optionally "check their work." Compared to participants who did not have the answer key, those with the answer key not only scored more highly, but also predicted (incorrectly) that they would score more highly on a subsequent test. One way to interpret this result is that participants had a strong prior belief in their ability, which led them to discard the auxiliary hypothesis that cheating aided their score, thereby inflating estimates of their own ability.

In settings where people might have some control over their observations, beliefs about rank or personal ability are closely connected to beliefs about controllability (Huys and Dayan, 2009). If a utility-maximizing agent believes that the world is controllable, then it is reasonable to assume that positive outcomes are more likely than negative outcomes, and hence negative outcomes are more likely to be explained away by alternative auxiliary hypotheses. For example, if you believe that you are a good test-taker (i.e., you have some control over test outcomes), then you may attribute poor test performance to the test difficulty rather than revising your beliefs about your own proficiency; this attribution is less plausible if you believe that you are a bad test-taker (i.e., you lack control over test outcomes). Thus, controllability is an important auxiliary hypothesis for interpreting feedback, with high perceived controllability leading to optimistic beliefs (Harris, 1996; Weinstein, 1980). The link between controllability and rank can be accommodated within the Bayesian framework, since we model the conditional distribution of the auxiliary hypothesis (controllability) given the central hypothesis (rank). This link is supported by studies showing that mood induction can bring about changes in beliefs about controllability (Alloy et al., 1981).

This analysis of controllability might provide insight into the psychopathology of asymmetric updating in response to positive and negative feedback. Individuals with depression do not show an optimism bias (so-called "depressive realism" Moore and Fresco, 2012), and Korn et al. (2014) demonstrated that this may arise from symmetric (unbiased) updating. One possibility is that this occurs because individuals with depression believe that the world is relatively uncontrollable—the key idea in the "learned helplessness" theory of depression (Abramson et al., 1978; Huys and Dayan, 2009; Seligman, 1975), which implies that they cannot take credit for positive outcomes any more than they can discount negative outcomes. Another possibility is that individuals with depression have a lower prior on rank, which would also lead to more symmetric updating compared to non-depressed individuals.

When placed in objectively uncontrollable situations, people will nonetheless perceive that they have control (Langer, 1975). According to the Bayesian analysis, this can arise when it is possible to discount unexpected outcomes in terms of an auxiliary hypothesis (e.g., fluke events, intrinsic variability, interventions by alternative causes) instead of reducing belief in control. As pointed out by Harris and Osman (2012), illusions of control typically arise in situations where cues indicate that controllability is plausible. For example, Langer (1975) showed that cues suggesting that one's opponent is incompetent inflate the illusion of control in a competitive setting, possibly by increasing the probability that the poor performance of the opponent is due to incompetence rather than the random nature of the outcomes. Another study showed that giving subjects an action that was in fact disconnected from the sequence of outcomes nonetheless inflated their perception that the sequence was controllable (Ladouceur and Sévigny, 2005). More generally, active involvement increases the illusion of control, as measured by the propensity for risk-taking: Davis et al. (2000) found that gamblers in real-world casinos placed higher bets on their own dice rolls than on others' dice rolls (see also Fernandez-Duque and Wifall, 2007; Gilovich and Douglas, 1986).

The basic lesson from all of these studies is that beliefs about controllability and rank can insulate an individual from the disconfirming effects of negative feedback. This response to negative feedback is rational under the assumption that alternative auxiliary hypotheses (e.g., statistical flukes) can absorb the blame.

The true self

Beliefs about the self provide a particularly powerful example of resistance to disconfirmation. People make a distinction between a "superficial" self and a "true" self, and these selves are associated with distinct patterns of behavior (Strohminger et al., 2017). In particular, people hold a strong prior belief that the true self is good (the central hypothesis h in our terminology). This proposition is supported by several lines of evidence. First, positive, desirable personality traits are viewed as more essential to the true self than negative, undesirable traits (Haslam et al., 2004). Second, people feel that they know someone most deeply when given positive information about them (Christy et al., 2017). Third, negative changes in traits are perceived as more disruptive to the true self than positive changes (De Freitas et al., 2017; Molouki and Bartels, 2017).

The key question for our purposes is what happens when one observes bad behavior: do people revise their belief in the goodness of the actor's true self? The answer is largely no. Bad behavior is attributed to the superficial self, whereas good behavior is attributed to the true self (Newman et al., 2014). This tendency is true even of individuals who generally have a negative attitude towards others, such as misanthropes and pessimists (De Freitas et al., 2016). And even if people are told explicitly that an actor's true self is bad, they are still reluctant to see the actor as truly bad (Newman et al., 2015). Conversely, observing positive changes in behavior (e.g., becoming an involved father after being a deadbeat) are perceived as indicating "self-discovery" (Bench et al., 2015; De Freitas et al., 2017).

These findings support the view that belief in the true good self shapes the perception of evidence about other individuals: evidence that disconfirms this belief tends to be discounted. The Bayesian framework suggests that this may occur because people infer alternative auxiliary hypotheses, such as situational factors that sever the link between the true self and observed behavior (e.g., he behaved badly because is mother just died). However, this possibility remains to be studied directly.

Stereotype updating

Stereotypes exert a powerful influence on our thinking about other people, but where do they come from? We are not born with strong beliefs about race, ethnicity, gender, religion, or sexual orientation; these beliefs must be learned from experience. What is remarkable is the degree to which stereotypes, once formed, are stubbornly resistant to updating (see Hewstone, 1994, for a review). As Lippmann (1922) remarked, "There is nothing so obdurate to education or criticism as the stereotype."

One possible explanation is that stereotypes are immunized from disconfirmation by flexible auxiliary hypotheses. This explanation fits well with the observation that individuals whose traits are inconsistent with a stereotype are segregated into "subtypes" without diluting the stereotype (Hewstone, 1994; Johnston and Hewstone, 1992; Weber and Crocker, 1983). For example, Weber and Crocker (1983) found that stereotypes were updated more when inconsistent traits were dispersed across multiple individuals rather than concentrated in a few individuals, consistent with the idea that concentration of inconsistencies licenses the auxiliary hypothesis that the individuals are outliers, and therefore do not reflect upon the group as a whole. An explicit sorting task supported this conclusion: inconsistent individuals tended to be sorted into separate groups (see also Johnston and Hewstone, 1992).

These findings have been simulated by a recurrent

connectionist model of stereotype judgment (Van Rooy et al., 2003). The key mechanism underlying subtyping is the competition between "group" units and "individual" units, such that stereotype-inconsistent information will be captured by individual units, provided the inconsistencies are concentrated in specific individuals. When the inconsistencies are dispersed, the group units take responsibility for them, updating the group stereotype accordingly. Another finding, also supported by connectionist modeling (Queller and Smith, 2002), is that individuals with moderate inconsistencies cause more updating than individuals with extreme inconsistencies. The logic is once again that extreme inconsistencies cause the individual to be segregated from the group stereotype.

Extinction learning

Like stereotypes, associative memories—in particular fear memories—are difficult to extinguish once formed. For example, in a typical fear conditioning paradigm, a rat is exposed to repeated tone-shock pairings; after only a few pairings, the rat will reliably freeze in response to the tone, indicating its anticipation of an upcoming shock. It may take dozens of tone-alone pairings to return the animal to its preconditioning response to the tone, indicating that extinction is much slower than acquisition. Importantly, the fact that the rat has returned to baseline does not mean that it has unlearned the fear memory. Under appropriate conditions, the rat's fear memory will return (Bouton, 2004). For example, simply waiting a month before testing the rat's response to the tone is sufficient to reveal the dormant fear, a phenomenon known as *spontaneous recovery* (Rescorla, 2004).

As with stereotype updating, one possibility is that conditioned fear is resistant to inconsistent information presented during extinction because the extinction trials are regarded as outliers or possibly subtypes. Thus, although fear can be temporarily reduced during extinction, it is not erased because the subtyping process effectively immunizes the fear memory from disconfirmation. In support of this view, there are suggestive parallels with stereotype updating. Analogous to the dispersed inconsistency conditions studied by Weber and Crocker (1983) and Johnston and Hewstone (1992), performing extinction in multiple contexts reduces the return of fear (Chelonis et al., 1999; Gunther et al., 1998). Analogous to the moderate versus extreme inconsistency manipulation (Queller and Smith, 2002), gradually reducing the frequency of tone-shock pairs during extinction prevents the return of fear (Gershman et al., 2013), possibly by titrating the size of the error signal driving memory updating (see also Gershman et al., 2014). More generally, it has been argued that effective memory updating procedures must control the magnitude of inconsistency between observations and the memory-based expectation, in order to prevent new memories from being formed to accommodate the inconsistent information (Gershman et al., 2017).

Conspiracy theories

As defined by Sunstein and Vermeule (2009), a conspiracy theory is "an effort to explain some event or practice by reference to the machinations of powerful people, who attempt to conceal their role (at least until their aims are accomplished)" (p. 205). Conspiracy theories are interesting from the perspective of auxiliary hypotheses because they often require a spiraling proliferation of auxiliaries to stay afloat. Each tenuous hypothesis needs an additional tenuous hypothesis to lend it plausibility, which in turn needs more tenuous hypotheses, until the theory embraces an enormous explanatory scope. For example, people who believe that the Holocaust was a hoax need to explain why the population of European Jews declined by 6 million during World War II; if they claim that the Jews immigrated to Israel and other countries, then they need to explain the discrepancy with immigration statistics, and if they claim that these statistics are false, then they need to explain why they were falsified, and so on.

Because conspiracy theories tend to have an elaborate support structure of auxiliary hypotheses, disconfirming evidence can be effectively explained away, commonly by undermining the validity of the evidence source. As Sunstein and Vermeule (2009) put it:

> Conspiracy theories often attribute extraordinary powers to certain agents—to plan, to control others, to maintain secrets, and so forth. Those who believe that those agents have such powers are especially unlikely to give respectful attention to debunkers, who may, after all, be agents or dupes of those who are responsible for the conspiracy in the first instance... The most direct governmental technique for dispelling false (and also harmful) beliefs—providing credible public information—does not work, in any straightforward way, for conspiracy theories. This extra resistance to correction through simple techniques is what makes conspiracy theories distinctively worrisome. (p. 207)

This description conforms to the Bayesian theory's prediction that a sparse, deterministic set of *ad hoc* auxil-

iary hypotheses can serve to explain away disconfirming data. In particular, conspiracy theorists use a large set of auxiliary hypotheses that perfectly (i.e., deterministically) predict the observed data and only the observed data (sparsity). This "drive for sense-making" (Chater and Loewenstein, 2016) is rational if the predictive power of a conspiracy theory outweighs the penalty for theory complexity—the Bayesian "Occam's razor" (MacKay, 2003).

Some evidence suggests that the tendency to endorse conspiracy theories is a personality trait or cognitive style: people who endorse one conspiracy theory tend to also endorse other conspiracy theories (Goertzel, 1994; Lewandowsky et al., 2013). One possibility is that this reflects parametric differences in probabilistic assumptions across individuals, such that people with very sparse and deterministic priors will be more likely to find conspiracy theories plausible.

Religious belief

While conspiracy theories are promulgated by relatively small groups of people, religious beliefs are shared by massive groups of people. Yet most of these people have little or no direct evidence for God: few have witnessed a miracle, spoken to God, or wrestled with an angel in their dreams. In fact, considerable evidence, at least on the surface, argues against belief in God, such as the existence of evil and the historical inaccuracy of the Bible.

One of the fundamental problems in the philosophy of religion is to understand the epistemological basis for religious beliefs—are they justified (Swinburne, 2004), or are they fictions created by psychological biases and cultural practices (Boyer, 2003)? Central to this debate is the status of evidence for the existence of God, such as reports of miracles. Following Hume (1748), a miracle is conventionally defined as "a transgression of a law of nature by a particular volition of the Deity"(p. 173). Hume famously argued that the evidence for miracles will always be outweighed by the evidence against them, since miracles are one-time transgressions of "natural" laws that have been established on the basis of countless observations. It would require unshakeable faith in the testimony of witnesses to believe in miracles, whereas in fact (Hume argues) testimony typically originates among uneducated, ignorant people.

As a number of philosophers (e.g., Earman, 2000; Swinburne, 1970) have pointed out, Hume's argument is weakened when one considers miracles through the lens of probability. Even if the reliability of individual witnesses was low, a sufficiently large number of such witnesses should be provide strong evidence for a miracle. Likewise, our beliefs about natural laws are based on a finite amount of evidence, possibly from sources of varying reliability, and hence are subject to the same probabilistic considerations. Whether or not the probabilistic analysis supports the existence of God depends on the amount and quality of evidence (both from experiment and hearsay) relative to the prior. Indeed, the same analysis has been used to deny the existence of God (Howson, 2011).

The probabilistic analysis of miracles provides another example of auxiliary hypotheses in action. The evidential impact of alleged miracles depends on auxiliary hypotheses about the reliability of testimony. If one is a religious believer, one can discount the debunking of miracles by questioning the evidence for natural laws. For example, some creationists argue that the fossil record is fake. Conversely, a non-believer can discount the evidence for miracles by questioning the eyewitness testimony, as Hume did. One retort to this view is that symmetry is misleading: the reliability of scientific evidence is much stronger than the historical testimony (e.g., Biblical sources). However, if one has a strong *a priori* belief in an omnipotent and frequently inscrutable God, then it may appear more plausible that apparent disconfirmations are simply examples of this inscrutability. In other words, if one believes in intelligent design, then scientific evidence that contradicts religious sources may be interpreted as evidence for our ignorance of the true design.[12]

Conceptual change in childhood

Children undergo dramatic restructuring of their knowledge during development, inspiring analogies with conceptual change in science (Carey, 2009; Gopnik, 2012). According to this "child-as-scientist" analogy, children engage in many of the same epistemic practices as scientists: probabilistically weighing evidence for different theories, balancing simplicity and fit, inferring causal relationships, carrying out experiments. If this analogy holds, then we should expect to see signs of resistance to disconfirmation early in development. In particular, Gopnik and Wellman (1992) have argued that children form *ad hoc* auxiliary hypotheses to reason about anomalous data until they can discover more coherent alternative theories.

[12] This point is closely related to the position known as *skeptical theism* (McBrayer, 2010), which argues that our inability to apprehend God's reasons for certain events (e.g., evil) does not justify the claim that no such reasons exist. This position undercuts inductive arguments against the existence of God that rely on the premise that no reasons exist for certain events.

For example, upon being told that the earth is round, some children preserve their preinstructional belief that the earth is flat by inferring that the earth is disc-shaped (Vosniadou and Brewer, 1992). After being shown two blocks of different weights hitting the ground at the same time when dropped from the same height, some middle-school students inferred that they hit the ground at different times but the difference was too small to observe, or that the blocks were in fact (contrary to the teacher's claims) the same weight (Champagne et al., 1985). Children who hold a geometric-center theory of balancing believe that blocks must be balanced in the middle; when faced with the failure of this theory applied to uneven blocks, children declare that the uneven blocks are impossible to balance (Karmiloff-Smith and Inhelder, 1975).

Experimental work by Schulz et al. (2008) has illuminated the role played by auxiliary hypotheses in children's causal learning. In these experiments, children viewed contact interactions between various blocks, resulting in particular outcomes (e.g., a train noise or a siren noise). Children then made inferences about novel blocks based on ambiguous evidence. The data suggest that children infer abstract laws that describe causal relations between classes of blocks. p (see also Saxe et al., 2005; Schulz and Sommerville, 2006). Schulz and colleagues argue for a connection between the rapid learning abilities of children (supported by abstract causal theories) and resistance to disconfirmation: the explanatory scope of abstract causal laws confer a strong inductive bias that enables learning from small amounts of data, and this same inductive bias confers robustness in the face of anomalous data by assigning responsibility to auxiliary hypotheses (e.g., hidden causes). A single anomaly will typically be insufficient to disconfirm an abstract causal theory that explains a wide range of data.

The use of auxiliary hypotheses has important implications for education. In their discussion of the educational literature, Chinn and Brewer (1993) point out that anomalous data are often used in the classroom to spur conceptual change, yet "the use of anomalous data is no panacea. Science students frequently react to anomalous data by discounting the data in some way, thus preserving their preinstructional theories" (p. 2). They provide examples of children employing a variety of discounting strategies, such as ignoring anomalous data, excluding it from the domain of the theory, holding it in abeyance (promising to deal with it later), and reinterpreting it. Careful attention to these strategies leads to pedagogical approaches that more effectively produce theory change. For example, Chinn and Brewer recommend helping children construct necessary background knowledge before introduction of the anomalous data, combined with the presentation of an intelligible and plausible alternative theory. In addition, bolstering the credibility of the anomalous data, avoiding ambiguities, and using multiple lines of evidence can be effective at producing theory change.

Is the Bayesian analysis falsifiable?

The previous sections have illustrated the impressive scope of the Bayesian analysis, but is it too impressive? Could it explain anything if we're creative enough at devising priors and auxiliaries that conform to the model's predictions? In other words, are Bayesians falling victim to their own Duhem-Quine thesis? Some psychologists say yes—that the success or failure of Bayesian models of cognition hinges on *ad hoc* choices of priors and likelihoods that conveniently fit the data (Bowers and Davis, 2012; Marcus and Davis, 2013).

It is true that Bayesian models can be abused in this way, and perhaps sometimes are. Nonetheless, Bayesian models *are* falsifiable, because their key predictions are not particular beliefs but particular regularities in belief updating. If I can independently measure (or experimentally impose) your prior and likelihood, then Bayes' rule dictates one and only one posterior. If this posterior does not conform to Bayes' rule, then the model has been falsified. Many tests of this sort of have been carried out, with the typical result (e.g., Evans et al., 2002) being that posterior judgments utilize both the prior and the likelihood, but do not precisely follow Bayes' rule (in some cases relying too much on the prior, and in other cases relying too much on the likelihood). The point here is not to establish whether people carry out exact Bayesian inference (they almost surely do not; see Dasgupta et al., 2017), but rather to show that they are not completely arbitrary.

As this article has emphasized, theories consist of multiple hypotheses (some central, some auxiliary) that work in concert to produce observations. Falsification of theories rests upon isolation and evaluation of these individual components; the theory as a whole cannot be directly falsified (Quine, 1951). The same is true for the Bayesian analysis of auxiliary hypotheses. In order to test this account, we would first need to independently establish the hypothesis space, the likelihood, and the prior. A systematic study of this sort has yet to be undertaken.

Conclusions

No one likes being wrong, but most of us believe that we *can* be wrong—that we would revise our be-

liefs when confronted with compelling disconfirmatory evidence. We conventionally think of our priors as inductive biases that may eventually be relinquished as we observe more data. However, priors also color our interpretation of data, determining how their evidential impact should be distributed across the web of beliefs. Certain kinds of probabilistic assumptions about the world lead one's beliefs (under perfect rationality) to be remarkably resistant to disconfirmation, in some cases even transmuting disconfirmation into confirmation. This should not be interpreted as an argument that people are perfectly rational, only that many aspects of their behavior that seem irrational on the surface may in fact be compatible with rationality when understood in terms of reasoning about auxiliary hypotheses.

An important implication is that if we want to change the beliefs of others, we need to attend to the structure of their belief systems rather than (or in addition to) the errors in their belief updating mechanisms. Rhetorical tactics such as exposing hypocrisies, staging interventions, declaiming righteous truths, and launching informational assaults against another person's central hypotheses are all doomed to be relatively ineffective from the point of view articulated here. To effectively persuade, one must incrementally chip away at the "protective belt" of auxiliary hypotheses until the central hypothesis can be wrested loose. The inherent laboriousness of this tactic may be why social and scientific progress is so slow, even with the most expert of persuasion artists.

Appendix: a sparse prior over auxiliary hypotheses

In this section, we define a sparse prior over auxiliary hypotheses using the Dirichlet distribution, which is the conjugate prior for the multinomial distribution. We focus on the case where the number of possible auxiliary hypotheses has a finite value (denoted by K), though extensions to infinite spaces are possible (Gershman and Blei, 2012). The symmetric Dirichlet probability density function over the K-simplex is given by:

$$P(\theta) = \frac{\Gamma(\alpha K)}{\Gamma(\alpha)^K} \prod_{k=1}^{K} \theta_k^{\alpha-1}, \tag{8}$$

where $\Gamma(\cdot)$ is the Gamma function, and $\alpha > 0$ is a *concentration parameter* that controls the sparsity of the distribution. As α approaches 0, the resulting distribution over auxiliary hypotheses, $P(a|\theta)$, places most of its probability mass on a small number of auxiliary hypotheses, whereas larger values of α induce distributions that evenly distribute their mass.

What are the consequences for the posterior over auxiliaries under the sparsity assumption (α close to 0)? Let us consider the case where auxiliary a_k predicts the observed data d with probability π_k (marginalizing over h). The posterior distribution is given by:

$$P(a_k|d) = \sum_i \frac{\alpha + \mathbb{I}[i = k]}{K\alpha + 1} \frac{\pi_i}{\sum_j \pi_j}, \tag{9}$$

where $\mathbb{I}[\cdot] = 1$ if its argument is true, 0 otherwise. In the sparse limit ($\alpha \to 0$), the posterior probability of an auxiliary is proportional to its agreement with the data: $P(a_k|d) \propto \pi_k$. If we restrict ourselves to auxiliaries that predict the data perfectly ($\pi_k = 1$) or not at all ($\pi_k = 0$), then the resulting posterior will be uniform over auxiliaries consistent with the data. It follows that $P(d|h\neg a) = 1$ in the sparse limit. Thus, sparsity favors auxiliaries that place high probability on the data, consistent with the assumptions underlying the analysis of Strevens (2001).

Acknowledgments

I am grateful to Michael Strevens, Josh Tenenbaum, Tomer Ullman, Alex Holcombe and Nick Chater for helpful discussions. This work was supported by the Center for Brains, Minds & Machines (CBMM), funded by NSF STC award CCF-1231216.

References

Abramson, L. Y., Seligman, M. E., and Teasdale, J. D. (1978). Learned helplessness in humans: Critique and reformulation. *Journal of Abnormal Psychology*, 87:49–74.

Alloy, L. B., Abramson, L. Y., and Viscusi, D. (1981). Induced mood and the illusion of control. *Journal of Personality and Social Psychology*, 41:1129–1140.

Austerweil, J. L. and Griffiths, T. L. (2011). Seeking confirmation is rational for deterministic hypotheses. *Cognitive Science*, 35:499–526.

Austerweil, J. L. and Griffiths, T. L. (2013). A nonparametric Bayesian framework for constructing flexible feature representations. *Psychological Review*, 120:817–851.

Bench, S. W., Schlegel, R. J., Davis, W. E., and Vess, M. (2015). Thinking about change in the self and others: The role of self-discovery metaphors and the true self. *Social Cognition*, 33:169–185.

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychological Review*, 94:115–147.

Bouton, M. E. (2004). Context and behavioral processes in extinction. *Learning & Memory*, 11:485–494.

Bovens, L. and Hartmann, S. (2002). Bayesian networks and the problem of unreliable instruments. *Philosophy of Science*, 69:29–72.

Bowers, J. S. and Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin*, 138:389–414.

Boyer, P. (2003). Religious thought and behaviour as by-products of brain function. *Trends in Cognitive Sciences*, 7:119–124.

Buchanan, D. W. and Sobel, D. M. (2011). Children posit hidden causes to explain causal variability. In *Proceedings of the 33rd annual conference of the cognitive science society*.

Buchanan, D. W., Tenenbaum, J. B., and Sobel, D. M. (2010). Edge replacement and nonindependence in causation. In *Proceedings of the 32nd annual conference of the cognitive science society*, pages 919–924.

Carey, S. (2009). *The Origin of Concepts*. Oxford University Press.

Champagne, A., Gunstone, R, F., and Klopfer, L, E. (1985). Instructional consequences of studentsÕ knowledge about physical phenomena. In West, L. and Pines, A., editors, *Cognitive structure and conceptual change*, pages 61–68. Academic Press.

Chance, Z., Norton, M. I., Gino, F., and Ariely, D. (2011). Temporal view of the costs and benefits of self-deception. *Proceedings of the National Academy of Sciences*, 108:15655–15659.

Chater, N. and Loewenstein, G. (2016). The under-appreciated drive for sense-making. *Journal of Economic Behavior & Organization*, 126:137–154.

Chelonis, J. J., Calton, J. L., Hart, J. A., and Schachtman, T. R. (1999). Attenuation of the renewal effect by extinction in multiple contexts. *Learning and Motivation*, 30:1–14.

Chinn, C. A. and Brewer, W. F. (1993). The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science instruction. *Review of Educational Research*, 63:1–49.

Christy, A. G., Kim, J., Vess, M., Schlegel, R. J., and Hicks, J. A. (2017). The reciprocal relationship between perceptions of moral goodness and knowledge of othersÕ true selves. *Social Psychological and Personality Science*.

Churchland, P. M. (1979). *Scientific realism and the plasticity of mind*. Cambridge University Press.

Cook, J. and Lewandowsky, S. (2016). Rational irrationality: Modeling climate change belief polarization using Bayesian networks. *Topics in Cognitive Science*, 8:160–179.

Dasgupta, I., Schulz, E., and Gershman, S. J. (2017). Where do hypotheses come from? *Cognitive Psychology*, 96:1–25.

Davis, D., Sundahl, I., and Lesbo, M. (2000). Illusory personal control as a determinant of bet size and type in casino craps games1. *Journal of Applied Social Psychology*, 30:1224–1242.

De Freitas, J., Sarkissian, H., Newman, G., Grossmann, I., De Brigard, F., Luco, A., and Knobe, J. (2016). Consistent belief in a good true self in misanthropes and three interdependent cultures. *Unpublished manuscript*.

De Freitas, J., Tobia, K. P., Newman, G. E., and Knobe, J. (2017). Normative judgments and individual essence. *Cognitive Science*, 41:382–402.

Dorling, J. (1979). Bayesian personalism, the methodology of scientific research programmes, and Duhem's problem. *Studies in History and Philosophy of Science Part A*, 10:177–187.

Duhem, P. M. (1954). *The Aim and Structure of Physical Theory*. Princeton University Press.

Earman, J. (1992). *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*. MIT Press.

Earman, J. (2000). *Hume's Abject Failure: The Argument Against Miracles*. Oxford University Press.

Eil, D. and Rao, J. M. (2011). The good news-bad news effect: asymmetric processing of objective information about yourself. *American Economic Journal: Microeconomics*, 3:114–138.

Evans, J. S. B., Handley, S. J., Over, D. E., and Perham, N. (2002). Background beliefs in Bayesian inference. *Memory & Cognition*, 30:179–190.

Fernandez-Duque, D. and Wifall, T. (2007). Actor/observer asymmetry in risky decision making. *Judgment and Decision Making*, 2:1.

Feyerabend, P. (1975). *Against Method*. Verso.

Fitelson, B. and Waterman, A. (2005). Bayesian confirmation and auxiliary hypotheses revisited: A reply to Strevens. *The British Journal for the Philosophy of Science*, 56:293–302.

Fodor, J. (1984). Observation reconsidered. *Philosophy of Science*, 51:23–43.

Frosch, C. A. and Johnson-Laird, P. N. (2011). Is everyday causation deterministic or probabilistic? *Acta Psychologica*, 137:280–291.

Garrett, N. and Sharot, T. (2017). Optimistic update bias holds firm: Three tests of robustness following shah et al. *Consciousness and Cognition*, 50:12–22.

Gershman, S. J. and Blei, D. M. (2012). A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*, 56:1–12.

Gershman, S. J., Jones, C. E., Norman, K. A., Monfils, M.-H., and Niv, Y. (2013). Gradual extinction prevents the return of fear: implications for the discovery of state. *Frontiers in Behavioral Neuroscience*, 7:164.

Gershman, S. J., Monfils, M.-H., Norman, K. A., and

Niv, Y. (2017). The computational nature of memory modification. *eLife*, 6:e23763.

Gershman, S. J., Radulescu, A., Norman, K. A., and Niv, Y. (2014). Statistical computations underlying the dynamics of memory updating. *PLoS Computational Biology*, 10:e1003939.

Gervais, S. and Odean, T. (2001). Learning to be overconfident. *Review of Financial studies*, 14:1–27.

Gibson, B. and Sanbonmatsu, D. M. (2004). Optimism, pessimism, and gambling: The downside of optimism. *Personality and Social Psychology Bulletin*, 30:149–160.

Gilovich, T. (1983). Biased evaluation and persistence in gambling. *Journal of Personality and Social Psychology*, 44:1110–1126.

Gilovich, T. (1991). *How We Know What Isn't So*. Simon and Schuster.

Gilovich, T. and Douglas, C. (1986). Biased evaluations of randomly determined gambling outcomes. *Journal of Experimental Social Psychology*, 22:228–241.

Gino, F., Norton, M. I., and Weber, R. A. (2016). Motivated Bayesians: Feeling moral while acting egoistically. *The Journal of Economic Perspectives*, 30:189–212.

Goertzel, T. (1994). Belief in conspiracy theories. *Political Psychology*, 15:731–742.

Gopnik, A. (2012). Scientific thinking in young children: Theoretical advances, empirical research, and policy implications. *Science*, 337:1623–1627.

Gopnik, A. and Wellman, H. M. (1992). Why the child's theory of mind really is a theory. *Mind & Language*, 7:145–171.

Gregory, R. L. (1970). *The Intelligent Eye*. Weidenfeld and Nicolson.

Griffiths, T. L. and Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review*, 116:661–716.

Grünbaum, A. (1962). The falsifiability of theories: total or partial? a contemporary evaluation of the Duhem-Quine thesis. *Synthese*, 14:17–34.

Gunther, L. M., Denniston, J. C., and Miller, R. R. (1998). Conducting exposure treatment in multiple contexts can prevent relapse. *Behaviour Research and Therapy*, 36:75–91.

Harding, S. (1976). *Can theories be refuted?: Essays on the Duhem-Quine thesis*. D. Reidel Publishing Company.

Harris, A. J. and Osman, M. (2012). The illusion of control: A bayesian perspective. *Synthese*, 189:29–38.

Harris, P. (1996). Sufficient grounds for optimism?: The relationship between perceived controllability and optimistic bias. *Journal of Social and Clinical Psychology*, 15:9–52.

Haslam, N., Bastian, B., and Bissett, M. (2004). Essentialist beliefs about personality and their implications. *Personality and Social Psychology Bulletin*, 30:1661–1673.

Helmholtz, H. v. (1867). *Handbuch der physiologischen Optik*. Voss.

Hempel, C. G. (1966). *Philosophy of Natural Science*. Prentice-Hall.

Hendrickson, A. T., Navarro, D. J., and Perfors, A. (2016). Sensitivity to hypothesis size during information search. *Decision*, 3:62–80.

Hewstone, M. (1994). Revision and change of stereotypic beliefs: In search of the elusive subtyping model. *European Review of Social Psychology*, 5:69–109.

Howson, C. (2011). *Objecting to God*. Cambridge University Press.

Howson, C. and Urbach, P. (2006). *Scientific reasoning: the Bayesian approach*. Open Court Publishing.

Hromádka, T., DeWeese, M. R., and Zador, A. M. (2008). Sparse representation of sounds in the unanesthetized auditory cortex. *PLoS Biology*, 6:e16.

Hume, D. (1748). *An Enquiry Concerning Human Understanding*.

Huys, Q. J. and Dayan, P. (2009). A Bayesian formulation of behavioral control. *Cognition*, 113:314–328.

Jaynes, E. T. (2003). *Probability Rheory: The Logic of Science*. Cambridge University Press.

Jern, A., Chang, K.-M. K., and Kemp, C. (2014). Belief polarization is not always irrational. *Psychological Review*, 121:206–224.

Johnston, L. and Hewstone, M. (1992). Cognitive models of stereotype change: 3. subtyping and the perceived typicality of disconfirming group members. *Journal of Experimental Social Psychology*, 28:360–386.

Jones, M. and Love, B. C. (2011). Bayesian fundamentalism or enlightenment? on the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences*, 34:169–188.

Karmiloff-Smith, A. and Inhelder, B. (1975). If you want to get ahead, get a theory. *Cognition*, 3:195–212.

Klayman, J. and Ha, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, 94:211–228.

Knill, D. C. and Richards, W. (1996). *Perception as Bayesian inference*. Cambridge University Press.

Koehler, J. J. (1993). The influence of prior beliefs on scientific judgments of evidence quality. *Organizational Behavior and Human Decision Processes*,

56:28–55.

Korn, C., Sharot, T., Walter, H., Heekeren, H., and Dolan, R. (2014). Depression is related to an absence of optimistically biased belief updating about future life events. *Psychological Medicine*, 44:579–592.

Korn, C. W., Prehn, K., Park, S. Q., Walter, H., and Heekeren, H. R. (2012). Positively biased processing of self-relevant social feedback. *Journal of Neuroscience*, 32:16832–16844.

Köszegi, B. (2006). Ego utility, overconfidence, and task choice. *Journal of the European Economic Association*, 4:673–707.

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. University of Chicago Press.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, 108:480–498.

Kuzmanovic, B. and Rigoux, L. (2017). Valence-dependent belief updating: Computational validation. *Frontiers in Psychology*, 8:1087.

Ladouceur, R. and Sévigny, S. (2005). Structural characteristics of video lotteries: Effects of a stopping device on illusion of control and gambling persistence. *Journal of Gambling Studies*, 21:117–131.

Lakatos, I. (1976). Falsification and the methodology of scientific research programmes. In *Can Theories be Refuted?*, pages 205–259. Springer.

Langer, E. J. (1975). The illusion of control. *Journal of personality and social psychology*, 32:311–328.

Laudan, L. (1990). Demystifying underdetermination. *Minnesota studies in the philosophy of science*, 14(1990):267–297.

Lefebvre, G., Lebreton, M., Meyniel, F., Bourgeois-Gironde, S., and Palminteri, S. (2017). Behavioural and neural characterization of optimistic reinforcement learning. *Nature Human Behaviour*, 1:0067.

Leiserowitz, A. A., Maibach, E. W., Roser-Renouf, C., Smith, N., and Dawson, E. (2013). Climategate, public opinion, and the loss of trust. *American behavioral scientist*, 57:818–837.

Lewandowsky, S., Oberauer, K., and Gignac, G. E. (2013). Nasa faked the moon landingNtherefore,(climate) science is a hoax: An anatomy of the motivated rejection of science. *Psychological Science*, 24:622–633.

Lippmann, W. (1922). *Public Opinion*. Harcourt, Brace.

Lord, C. G., Ross, L., and Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, 37:2098–2109.

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

MacKay, D. J. (2003). *Information Theory, Inference and Learning Algorithms*. Cambridge university press.

Marcus, G. F. and Davis, E. (2013). How robust are probabilistic models of higher-level cognition? *Psychological Science*, 24:2351–2360.

Mayo, D. G. (1997). Duhem's problem, the Bayesian way, and error statistics, or "what's belief got to do with it?". *Philosophy of Science*, 64:222–244.

Mayrhofer, R. and Waldmann, M. R. (2015). Sufficiency and necessity assumptions in causal structure induction. *Cognitive Science*, 40:2137–2150.

McBrayer, J. P. (2010). Skeptical theism. *Philosophy Compass*, 5:611–623.

McKenzie, C. R., Ferreira, V. S., Mikkelsen, L. A., McDermott, K. J., and Skrable, R. P. (2001). Do conditional hypotheses target rare events? *Organizational Behavior and Human Decision Processes*, 85:291–309.

Mckenzie, C. R. and Mikkelsen, L. A. (2000). The psychological side of Hempel's paradox of confirmation. *Psychonomic Bulletin & Review*, 7:360–366.

McKenzie, C. R. and Mikkelsen, L. A. (2007). A Bayesian view of covariation assessment. *Cognitive Psychology*, 54:33–61.

Molouki, S. and Bartels, D. M. (2017). Personal change and the continuity of the self. *Cognitive Psychology*, 93:1–17.

Moore, M. T. and Fresco, D. M. (2012). Depressive realism: A meta-analytic review. *Clinical Psychology Review*, 32:496–509.

Muentener, P. and Schulz, L. (2014). Toddlers infer unobserved causes for spontaneous events. *Frontiers in Psychology*, 5.

Navarro, D. J. and Perfors, A. F. (2011). Hypothesis generation, sparse categories, and the positive test strategy. *Psychological Review*, 118:120–134.

Newman, G. E., Bloom, P., and Knobe, J. (2014). Value judgments and the true self. *Personality and Social Psychology Bulletin*, 40:203–216.

Newman, G. E., De Freitas, J., and Knobe, J. (2015). Beliefs about the true self explain asymmetries based on moral judgment. *Cognitive Science*, 39:96–125.

Nisbett, R. E. and Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Prentice-Hall.

Oaksford, M. and Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, 101:608–631.

Olshausen, B. A. and Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning

a sparse code for natural images. *Nature*, 381:607–609.

Perfors, A. and Navarro, D. (2009). Confirmation bias is rational when hypotheses are sparse. *Proceedings of the 31st Annual Conference of the Cognitive Science Society*.

Poo, C. and Isaacson, J. S. (2009). Odor representations in olfactory cortex:"sparse" coding, global inhibition, and oscillations. *Neuron*, 62:850–861.

Popper, K. (1959). *The Logic of Scientific Discovery*. Harper & Row.

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

Queller, S. and Smith, E. R. (2002). Subtyping versus bookkeeping in stereotype learning and change: Connectionist simulations and empirical findings. *Journal of Personality and Social Psychology*, 82:300–313.

Quine, W. V. (1951). Two dogmas of empiricism. *The Philosophical Review*, pages 20–43.

Rescorla, R. A. (2004). Spontaneous recovery. *Learning & Memory*, 11:501–509.

Rock, I. (1983). *The Logic of Perception*. MIT Press.

Rosch, E. (1978). Principles of categorization. In *Cognition and Categorization*, pages 27–48. Hillsdale, NJ: Lawrence Erbaum Associates.

Saxe, R., Tenenbaum, J., and Carey, S. (2005). Secret agents inferences about hidden causes by 10- and 12-month-old infants. *Psychological Science*, 16:995–1001.

Schulz, L. E., Goodman, N. D., Tenenbaum, J. B., and Jenkins, A. C. (2008). Going beyond the evidence: Abstract laws and preschoolersÕ responses to anomalous data. *Cognition*, 109:211–223.

Schulz, L. E. and Sommerville, J. (2006). God does not play dice: Causal determinism and preschoolers' causal inferences. *Child Development*, 77:427–442.

Seligman, M. E. (1975). *Helplessness: On depression, development, and death*. WH Freeman/Times Books/Henry Holt & Co.

Shah, P., Harris, A. J., Bird, G., Catmur, C., and Hahn, U. (2016). A pessimistic view of optimistic belief updating. *Cognitive Psychology*, 90:71–127.

Sharot, T. (2011). *The Optimism Bias*. Vintage.

Sharot, T. and Garrett, N. (2016). Forming beliefs: why valence matters. *Trends in Cognitive Sciences*, 20:25–33.

Sharot, T., Korn, C. W., and Dolan, R. J. (2011). How unrealistic optimism is maintained in the face of reality. *Nature Neuroscience*, 14:1475–1479.

Stankevicius, A., Huys, Q. J., Kalra, A., and Seriès, P. (2014). Optimism as a prior belief about the probability of future reward. *PLoS Computational Biology*, 10:e1003605.

Stratton, G. M. (1897). Vision without inversion of the retinal image. *Psychological Review*, 4:341–360.

Strevens, M. (2001). The Bayesian treatment of auxiliary hypotheses. *The British Journal for the Philosophy of Science*, 52:515–537.

Strohminger, N., Knobe, J., and Newman, G. (2017). The true self: A psychological concept distinct from the self. *Perspectives on Psychological Science*.

Sunstein, C. R. and Vermeule, A. (2009). Conspiracy theories: Causes and cures. *Journal of Political Philosophy*, 17:202–227.

Swinburne, R. G. (1970). *The Concept of Miracle*. Springer.

Swinburne, R. G. (2004). *The Existence of God*. Oxford University Press.

Van Rooy, D., Van Overwalle, F., Vanhoomissen, T., Labiouse, C., and French, R. (2003). A recurrent connectionist model of group biases. *Psychological Review*, 110:536–563.

Vosniadou, S. and Brewer, W. F. (1992). Mental models of the earth: A study of conceptual change in childhood. *Cognitive Psychology*, 24:535–585.

Weber, R. and Crocker, J. (1983). Cognitive processes in the revision of stereotypic beliefs. *Journal of Personality and Social Psychology*, 45:961–977.

Weinstein, N. D. (1980). Unrealistic optimism about future life events. *Journal of personality and social psychology*, 39:806–820.

Wu, Y., Muentener, P., and Schulz, L. E. (2015). The invisible hand: toddlers connect probabilistic events with agentive causes. *Cognitive Science*, 40:1854–1876.

Yeung, S. and Griffiths, T. L. (2015). Identifying expectations about the strength of causal relationships. *Cognitive Psychology*, 76:1–29.