Hierarchical Inference as a Source of Human Biases

Paul B. Sharp[1,2,†], Isaac Fradkin[3,4,†], Eran Eldar[1,2]

[1] Department of Psychology, Hebrew University of Jerusalem, Jerusalem 9190501, Israel

[2] Department of Cognitive and Brain Sciences, Hebrew University of Jerusalem, Jerusalem 9190501, Israel

[3] Max Planck University College London Centre for Computational Psychiatry and Ageing Research, London WC1B 5EH, United Kingdom

[4] Wellcome Trust Centre for Neuroimaging, University College London, London WC1N 3BG, United Kingdom

† Equal contributors

Corresponding author: Eran Eldar,

Department of Psychology & Department of Cognitive and Brain Sciences,

Hebrew University of Jerusalem,

Jerusalem

Email: eran.eldar@mail.huji.ac.il

Abstract

The finding that human decision-making is systematically biased continues to have an immense impact on both research and policymaking. Prevailing views ascribe biases to limited computational resources, which require humans to resort to less costly resource-rational heuristics. Here, we propose that many biases in fact arise due to a computationally costly way of coping with uncertainty – namely, hierarchical inference – which by nature incorporates information that can seem irrelevant. We thus show how, in uncertain situations, Bayesian inference may avail of the environment's hierarchical structure, so as to reduce uncertainty at the cost of introducing bias. We illustrate how this account can explain a range of familiar biases, focusing in detail on the halo effect and on the neglect of base rates. In each case, we show how a hierarchical-inference account takes the characterization of a bias beyond phenomenological description by revealing the computations and assumptions it might reflect. Furthermore, we highlight new predictions entailed by our account concerning factors that could mitigate or exacerbate bias, some of which have already garnered empirical support. We conclude that a hierarchical inference account may inform scientists and policy makers with a richer understanding of the adaptive and maladaptive aspects of human decision-making.

Introduction

One of the most influential ideas in the study of human decision-making is that many of our intuitive decisions are based on simplifying heuristics that lead to systematic biases (Gilovich, Griffin, and Kahneman, 2002; **Box 1**). Instead of fully and properly considering relevant information, it is thought that we resort to heuristics due to the limited nature of our cognitive resources and a consequent need to minimize computational demands (Gilovich et al., 2002; Kahneman, 2011; Lieder, Griffiths, and Hsu, 2018; Lieder, Griffiths, M Huys, and Goodman, 2018). The far-reaching impact of this idea is highlighted by two Nobel Prizes in Economics awarded in 2002 and 2017 to researchers who developed it (Grüne-Yanoff, 2017; Guomei and Qicheng, 2003), and by its widespread influence on current social and economic policies (John, 2018; Schmidt, 2017; Schwartz, 2015; Thaler, 2018b, 2018a). Here, we propose that some of the most fundamental human biases do not reflect a computational limitation, but rather the faithful operation of an advanced form of inference, namely, hierarchical inference.

Box 1. What is a bias?

Biases are commonly thought of as inherently irrational influences of irrelevant information on people's judgments. For example, expecting good looking people to be more intelligent. In statistics, however, bias is defined slightly differently – it refers to a systematic deviation of inferred values from true values. Such statistical bias is in some cases rational since biased estimates can be more accurate than unbiased ones. For example, the way our perceptions are biased by our prior expectations can make our perceptions more accurate, provided that our prior expectations are well calibrated to the true probabilities of different percepts. In this paper, we show that some of the biases that have typically been thought of as irrational may in fact constitute rational statistical biases.

Contemporary cognitive science has given rise to a view of the human brain as a Bayesian inference machine (Friston, 2012; Griffiths,et al., 2010; Knill and Pouget, 2004; Piray and Daw, 2020; Summerfield and Tsetsos, 2012; Tenenbaum, Kemp, Griffiths, and Goodman, 2011). Examining this research reveals a key feature that distinguishes more recent models of rational inference from those that have been prevalent in the heuristics and biases literature: hierarchical structure (Benrimoh et al., 2018; Diaconescu et al., 2020;

Fradkin et al., 2020; Fradkin, Ludwig, Eldar, and Huppert, 2020; Glaze et al., 2018; Hesp et al., 2021; Lawson, Mathys, and Rees, 2017; Lee and Newell, 2011; Powers, Mathys, and Corlett, 2017; Qiu, Luu, and Stocker, 2020; Reed et al., 2020; Schustek, Hyafil, and Moreno-Bote, 2019; Siegel, Mathys, Rutledge, and Crockett, 2018; Smith, Thayer, Khalsa, and Lane, 2017; van Ravenzwaaij, Moore, Lee, and Newell, 2014). The appeal of hierarchical models is that they use the temporal and structural dependencies existing in the world around us to mitigate uncertainty (**Box 2**). This is achieved by informing inferences about variables of interest (e.g., the expected harvest of fruit from a specific tree) not only with observations that directly reflect the variables (e.g., fruit previously harvested from the tree), as simple inference would, but also with observations reflecting indirectly related variables (e.g., fruit harvested from other trees in the same valley). The result is more informed, and thus more precise (i.e., less uncertain), inferences. Ample evidence supports humans' pervasive use of such hierarchical inference in a range of cognitive functions, including perception (de Lange, Heilbron and Kok, 2018), social cognition (Gweon, 2021), and reinforcement learning (Behrens et al., 2008).

In this paper, we re-analyze past findings from the heuristics and biases literature to illustrate how the use of hierarchical inference can produce multiple classical decision biases. This form of inference, however, is costly to implement. How then can we reconcile evidence that people intuitively perform hierarchical inference with observations that people fail to do even simple Bayesian inference properly (Tversky and Kahneman, 1981a)? We propose that this apparent contradiction can be resolved by realizing that the use of indirectly relevant information to reduce uncertainty produces behaviors that only appear erroneous if we assume that people are attempting simple inference. Finally, we highlight unique predictions regarding how diminished neurocognitive resources and effort can be expected to mitigate,

rather than augment, biases, which sharply distinguishes our account of decision biases from previous accounts.

Box 2. What is hierarchical inference?

Statistical inference is deemed hierarchical when inferred variables occupy multiple levels in the inference model. Though hierarchical inference has many forms and uses (e.g., frequentist multilevel modelling), here we focus on hierarchical Bayesian inference as a model of cognitive function.

Consider an animal that wants to estimate which trees bear the most fruit. To do so, the animal chooses a tree based on a prior expectation concerning how fecund the tree may be, $\mathrm{p}(f_{\mathrm{tree1}})$. It then observes how much fruit there currently is on this tree, r_{tree1}, and finally updates its expectation concerning the tree's fecundity in light of the amount of fruit it just observed, $\mathrm{p}(f_{\mathrm{tree1}}|r_{\mathrm{tree1}})$. This updated expectation is referred to as a posterior probability, and is derived using simple Bayesian inference:

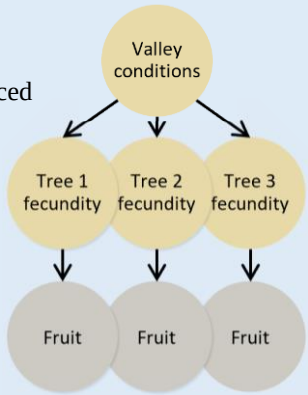
$$\mathrm{p}(f_{\mathrm{tree1}}|r_{\mathrm{tree1}}) \propto \mathrm{p}(r_{\mathrm{tree1}}|f_{\mathrm{tree1}})\mathrm{p}(f_{\mathrm{tree1}}).$$

This type of Bayesian inference, although itself complex, ignores how nearby trees from the same valley may be similarly fecund. Hierarchical inference uses such nested relationships to make more informed, and thus less uncertain, inferences: food from a tree depends on the qualities of that specific tree, $\mathrm{p}(r_{\mathrm{tree1}}|f_{\mathrm{tree1}})$, which themselves depend on qualities of the larger valley in which the tree grows, $\mathrm{p}(f_{\mathrm{tree1}}|f_{\mathrm{valley}})$, about which the animal may have some prior expectation, $\mathrm{p}(f_{\mathrm{valley}})$. After sampling a tree, the animal can use hierarchical inference to simultaneously update its interdependent expectations concerning the tree and the valley:

$$\mathrm{p}(f_{\mathrm{tree1}}, f_{\mathrm{valley}}|r_{\mathrm{tree1}}) \propto \mathrm{p}(r_{\mathrm{tree1}}|f_{\mathrm{tree1}})\mathrm{p}(f_{\mathrm{tree1}}|f_{\mathrm{valley}})\mathrm{p}(f_{\mathrm{valley}}).$$

The expected valley fecundity will thus come to reflect all trees that have been sampled in that valley. Critically, since expectations from trees are now influenced by expectations from the valley in which they reside, the animal's expectations from each tree will be *biased* by how much fruit it obtained from neighboring trees.

As depicted on the right, we can represent the nested relations involved in hierarchical inference using a graphical model. In the hierarchical models presented in this paper, variables of the model (gold circles) are estimated based on observations that lie at the bottom of the hierarchy (gray circles) and prior expectations concerning the top-level variables (not shown; often referred to as hyper-priors).



Hierarchical inference is intuitive for humans

Ample empirical evidence supports our use of hierarchical inference across a range of domains. Perceptual neuroscience, for example, has shown how higher-level inference about the general context portrayed by a visual stimulus (e.g., "I believe I'm looking at a picture of a typical day in a city") shapes lower-level inferences about individual objects within this context (e.g., "this blurry image must be a car parked next to a sidewalk"; de Lange, Heilbron and Kok, 2018). How we learn and make decisions has similarly been shown to involve higher-level inferences that guide lower-level inferences. For instance, inferences of rates of

change in reward or punishment guide inferences concerning the reward and punishment associated with specific choices (Behrens et al., 2008; de Berker et al., 2015; Eldar et al., 2016). Hierarchical inference is also prevalent in language. Even when we are very young, we routinely infer latent causes from the speech we hear, including goals and emotions that are not directly observable from others' speech, and these higher-level inferences guide our interpretation of subsequent speech (Gweon, 2021). Without this ability to infer hierarchical causes, we would misinterpret sarcasm for offense, humor for stupidity, and so on. Indeed, our behavioral flexibility is often predicated on employing hierarchical inference. It is thus unsurprising that hierarchical inference is implemented in leading frameworks of brain function, including predictive coding (Friston, 2012) and reinforcement learning (Bartolo and Averbeck, 2021).

In what follows, we illustrate how the operation of hierarchical inference could explain several well-established decision biases. In each case, we explore the novel insights that a hierarchical inference account affords. Importantly, our account does not seek to overturn all biases as non-biases, but rather pinpoint the computations that underly them and thus help determine whether, or in what circumstances, we can construe a bias as rational and desired.

The halo effect: a paradigmatic example of intuitive hierarchical inference

Consider, for instance, the well-known bias that goes by the name of 'halo effect' (Thorndike, 1920). A classic example of a halo effect manifests in grading an exam consisting of two open-ended questions: the evaluation of the question graded first can bias the evaluation of the second question. Thus, an exceptional first question makes it likely that the second question will be evaluated more highly than it otherwise would have been. This

behavior seems unacceptable because it leads to a different overall grade depending on the arbitrary factor of which question is graded first.

Such behavior, however, is natural under a hierarchical model, wherein the student's knowledge on individual questions is assumed to reflect their typical level of knowledge on the topic (**Figure 1a**, right). Under this assumption, given that the information a grader garners from a student's answer is a noisy representation of the student's true level of knowledge on the question, the grader's grade will more accurately reflect the student's knowledge if it forms a compromise between the information garnered from the answer and what is known about the student's typical level of knowledge. This typical level of knowledge (and the consistency of knowledge across questions) can be sequentially estimated from the student's answers to previous questions.

To illustrate how sequential hierarchical inference has the consequence that the order in which answers are graded impacts the overall grade, imagine that a student's answer to the first open-ended question (Q1) seems highly accurate (**Figure 1b**, green vertical line). Based on this impression and no specific prior expectation, the grader infers both the student knowledge for Q1 (**Figure 1b**, middle) and her general knowledge on the exam topic (**Figure 1b**, left). The latter estimate now provides a more precise prior expectation with regards to the student's knowledge on Q2 (**Figure 1b**, right plot, green). Consequently, the grading of Q2 is both more certain and pulled upwards relative to how it would have been had it also been graded without an informative prior expectation (**Figure 1b**, right plot, compare solid and dashed brown lines). Conversely, had the grader begun grading with Q2, her estimate of student knowledge would have been lower at the time she graded Q1. In this alternative scenario, Q2's grade would be unbiased whereas Q1's grade would be biased downwards. Sequential hierarchical inference thus explains how order effects may emerge in grading as an outcome of a rational process.

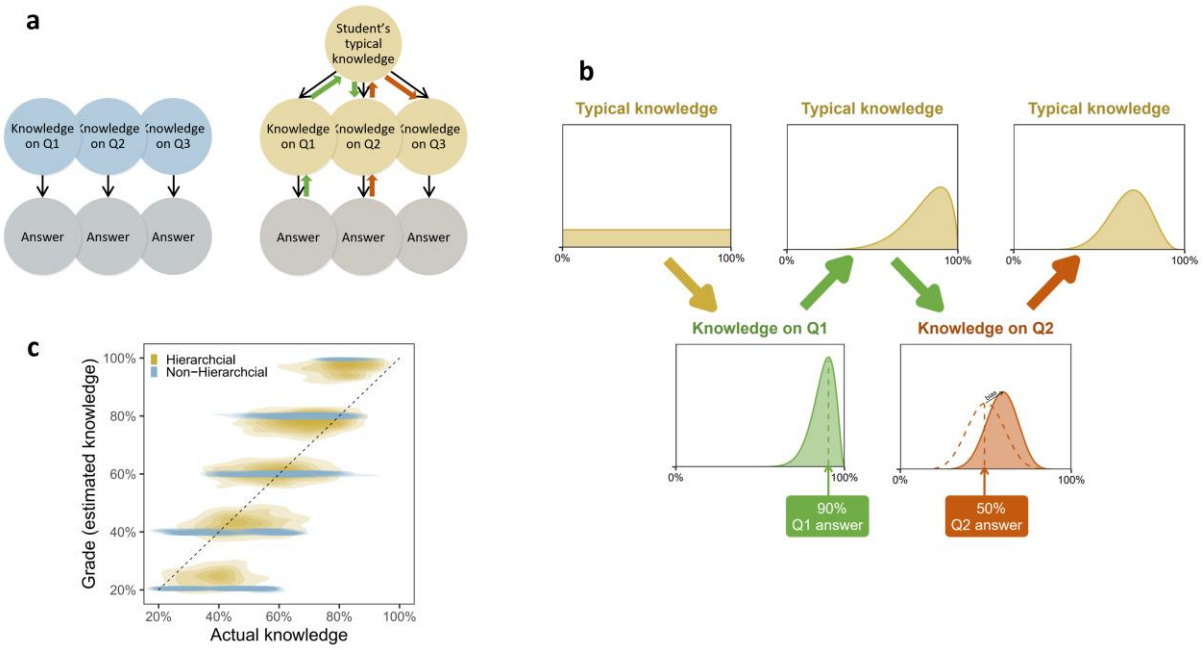


Figure 1. Hierarchical inference produces a halo effect in grading questions. (**a**) A non-hierarchical model (blue) assumes that knowledge on different questions is unrelated. In contrast, a hierarchical model (gold) assumes knowledge on different questions is linked via the student's typical level of knowledge in the subject matter (black arrows). It is thus sensible that the evaluation of one answer will inform the evaluation of a subsequent answer. This flow of information is illustrated by added green and brown arrows. Because the grader does not re-grade earlier answers after additional information about the student is garnered from later answers, earlier grades do not benefit from (i.e., are not biased by) this latter information. Such sequential inference is encountered in many real-world situations where choices must be enacted as evidence accrues. (**b**) Hierarchical inference concerning a student's typical knowledge (top row, gold), and specific knowledge on two questions, the first of which received a fairly accurate answer (Q1, green) whereas the second received a mediocre answer (Q2, brown). The grader begins with an assumption that the student's typical level of knowledge could equally be any level between 0% and 100% (illustrated in the top left plot by a uniform distribution). Inferred Q1 knowledge is thus unbiased, but it informs the grader's inference concerning the student's typical knowledge, which serves as a prior expectation for inferring Q2 knowledge. As a result, Q2 knowledge is inferred with higher certainty and an upwards bias compared to how it would have been inferred without hierarchical inference (dashed line). (**c**) Grading of individual questions by non-hierarchical (blue clouds) and hierarchical (gold clouds) inference as a function of the student's knowledge on the question, for five levels of seeming answer accuracy (20%, 40%, 60%, 80%, 100%) from a student whose typical level of knowledge lies in the range of 50% to 90%. Both types of grades aim to estimate the student's knowledge on each question. In grading a question, non-hierarchical inference only relies on the answer to that question. Thus, its grades precisely equal the answer's seeming accuracy. By contrast, hierarchical inference is informed by the levels of accuracy the student demonstrated in previous questions. Consequently, its grades smooth out the noise embedded in raw answers, and thus more faithfully reflect the student's knowledge on each question (i.e., the gold clouds are closer to the diagonal).

This analysis of the halo effect illustrates how, by accounting for dependencies between different sets of observations, hierarchical inference offers more accurate estimates and greater certainty about them. Though certainty has not been empirically investigated in this context, it is interesting to note, anecdotally, what happened when Daniel Kahneman shuffled students' exams to stop himself from being influenced by inferences about each

student's typical level of knowledge. "I was now less happy with and less confident in my grades", he wrote (Kahneman, 2011).

Hierarchical inference in other heuristics and biases

The relevance of hierarchical inference extends to a variety of established heuristics and biases that characterize human decision-making (**Figure 2**). For example, the impact of an incidental affective state on the evaluation of outcomes is typically regarded as an affective bias (Slovic, Finucane, Peters, and MacGregor, 2007). A hint that this bias may serve some form of inference is provided by the finding that the bias is mitigated if the affective state can be attributed to an unrelated cause (Schwarz and Clore, 1983). However, until recently, it remained unclear why by default people's judgments are influenced by non-specific affective states. More recent analysis has offered an answer to this conundrum, by showing that affective states may reflect hierarchical inference about general environmental changes that simultaneously increase (or decrease) the value of multiple related actions (e.g.,

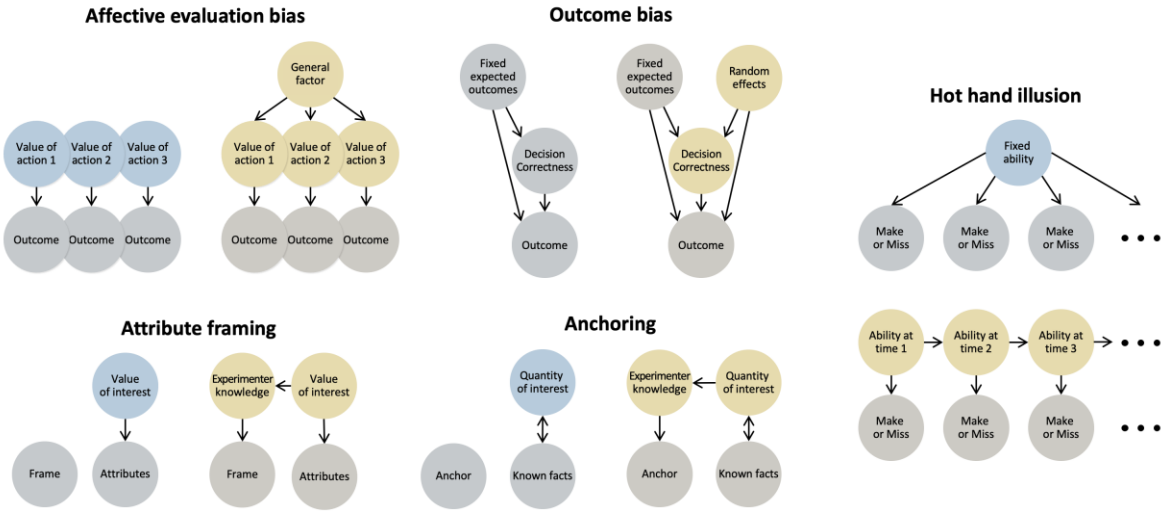


Figure 2. Biases that violate simple Bayesian inference but are consistent with hierarchical inference. Each bias constitutes a well-established property of human decision-making. From the point of view of non-hierarchical Bayesian inference (blue models) the biases are unjustified, but hierarchical models (gold) show how, given certain assumptions, the biases can reflect rational inference. Note that the gold 'hot-hand illusion' model is not strictly hierarchical, but it involves 'hierarchical-like' inference wherein inferred variables are constrained by other inferred variables. Bidirectional arrows are used to indicate that known facts may either depend on the quantity of interest or the quantity of interest may depend on them. See **Box 3** for equations describing each model.

a seasonal change that makes it easier to obtain food, water, and shelter). Affective biases may thus serve to properly correct learning concerning individual actions to account for general environmental changes (Eldar, Rutledge, Dolan, and Niv, 2016).

The effects of 'anchors' (Tversky and Kahneman, 1974) and 'frames' (Tversky and Kahneman, 1981b) on people's estimations and evaluations have similarly been regarded as irrational biases. Such biases manifest, for instance, when people give different answers to two logically equivalent but differently framed questions. Endeavoring to understand the root of this irrational behavior, substantial research has investigated the processes through which these biases are produced. Such research has found, for instance, that an anchoring bias may arise either because the anchor serves as an initial estimate that is gradually adjusted until reaching a plausible value (Tversky and Kahneman, 1974), or because the anchor primes relevant knowledge with which it is consistent (Strack and Mussweiler, 1997).

Elucidating the algorithm that produces a bias, however, does not necessarily reveal why a bias exists. Indeed, dissenting voices have suggested that anchors and frames should not be construed as exerting irrational influences since they may reflect relevant knowledge on the part of the individual who designed the decision problem. For instance, it has been suggested that "a speaker is likely making an unspoken recommendation when using a positive frame." (Gigerenzer, 2018; Sher and McKenzie, 2007). Such an inference may give rise to multiple different types of framing effects (Levin, Schneider, and Gaeth, 1998), since a speaker's recommendation could shed light on the value of an item (attribute framing) or indicate what type of outcomes (risky-choice framing) or features (goal framing) should be given priority. In agreement with this suggestion, more recent research has found that framing effects are eliminated when the frame is made uninformative by disambiguating provided information (Mandel, 2014), and anchoring biases are decreased or eliminated when anchors are made irrelevant (Fudenberg, Levine, and Maniadis, 2012; Ioannidis, Offerman,

and Sloof, 2020) or unnecessary (Jacowitz and Kahneman, 1995; Wilson et al., 1996). Accordingly, we propose that framing and anchoring biases constitute additional manifestations of hierarchical inference, wherein the frame or the anchor are used to infer relevant knowledge on the part of the experimenter, which is in turn used to infer the target quantity (**Figure 2**). Such hierarchical inference could be implemented by means of any of the algorithms previously suggested to produce these biases.

Finally, it is noteworthy that a previously established bias, the hot-hand fallacy in basketball (Gilovich, Vallone, and Tversky, 1985), has recently been shown to not be a fallacy (Miller and Sanjurjo, 2018; Ritzwoller and Romano, 2022), indicating that basketball viewers and players may be making well-founded 'hierarchical-like' inferences (**Figure 2**) when they identify a hot streak.

These biases offer an illustrative set of examples for how hierarchical inference may give rise to judgments that are biased yet rational. This is not to say that all biases can be explained in this way, nor that other biases that we have not discussed cannot. Thus, for instance, hierarchical inference might also give rise to the availability heuristic (Tversky and Kahneman, 1983), which may possibly be conceptualized as an inference of the frequency of an event based on how frequently we have previously encountered it, which is in turn inferred from how quickly information about it comes to mind (since repetition improves recall; Hintzman, 1976). To facilitate further investigation and quantitative testing of a hierarchical inference account of these and other biases, in **Box 3** we provide equations for generative hierarchical models that may explain the computations that produce each bias.

A hierarchical Bayesian interpretation of base rate neglect

The biases discussed so far demonstrate how rational hierarchical inference can lead to decisions that seem biased. However, these biases were never specifically perceived as

incompatible with the view of humans as Bayesian. In the next example, we use a hierarchical model to re-interpret a well-established bias that cast doubt on the ability of humans to perform Bayesian inference – base rate neglect (Tversky and Kahneman, 1981a). In a classic example illustrating base rate neglect, participants are asked to judge the probability that a car involved in an accident belonged to the blue cab company vs. the green cab company, given that only 15% of cabs are blue and that a moderately-reliable (i.e., 80% accurate) eyewitness said they saw a blue cab. The classic finding is that participants tend to underweight the 15% base rate, and thus overestimate the probability that the cab in the accident was blue. Interestingly, however, if participants are told instead that 15% of cabs *involved in accidents* are blue, base rate neglect is substantially diminished or even entirely absent (Bar-Hillel, 1980; Tversky and Kahneman, 1980). This difference has led the former type of base rate to be termed an *incidental* base rate, and the latter a *causal* base rate.

The hierarchical inference perspective offers a way to explain why people weight causal base rates more strongly than incidental base rates. Given a causal base rate, the probability that the cab in the accident was blue should be computed using a straightforward application of Bayes rule (**Figure 3a**, blue). This precise computation was previously used to derive the same optimal answer for causal and incidental base rates (Tversky and Kahneman, 1981a). However, a closer examination of the incidental base rate case suggests that it requires a more complex computation. This is because the proportion of blue cabs out of all cabs involved in accidents is determined not only by the proportion of blue cabs out of all cabs, but also by the relative proneness to accidents of cabs from the green and blue cab companies (**Figure 3a**, gold). Differences in accident proneness are likely, for instance, if the two cab companies operate in different areas, or if their driver hiring practices differ.

Though we are given no information about the relative proneness to accidents of the two cab companies, simply by accounting for our uncertainty about accident proneness,

hierarchical inference produces a different answer concerning the cab involved in the accident. To see this, consider that the proportion of blue cabs involved in accidents (i.e., causal base rate C) can be computed by multiplying the proportion of blue cabs (i.e., incidental base rate I) by their accident proneness (P):

$$C_{\text{blue}} = \frac{I_{\text{blue}}P_{\text{blue}}}{I_{\text{blue}}P_{\text{blue}} + I_{\text{green}}P_{\text{green}}}$$

The precise result of this computation depends on our prior assumption about accident proneness. If we assume that the accident proneness of the two companies is equal, then the causal and incidental base rates are identical, and therefore hierarchical and non-hierarchical inferences produce the same result (**Figure 3b**, left panel). By contrast, if we assume that accident proneness differs to an extreme extent such that one or the other company is responsible for 100% of the accidents, then the causal base rate (C_{blue}) is either 1 or 0 with equal probability, and thus, the posterior probability that the cab was blue matches the reliability of the witness (**Figure 3b**, right panel). That is, in this case the incidental base rate should be deemed completely irrelevant. Of course, a more reasonable assumption is that

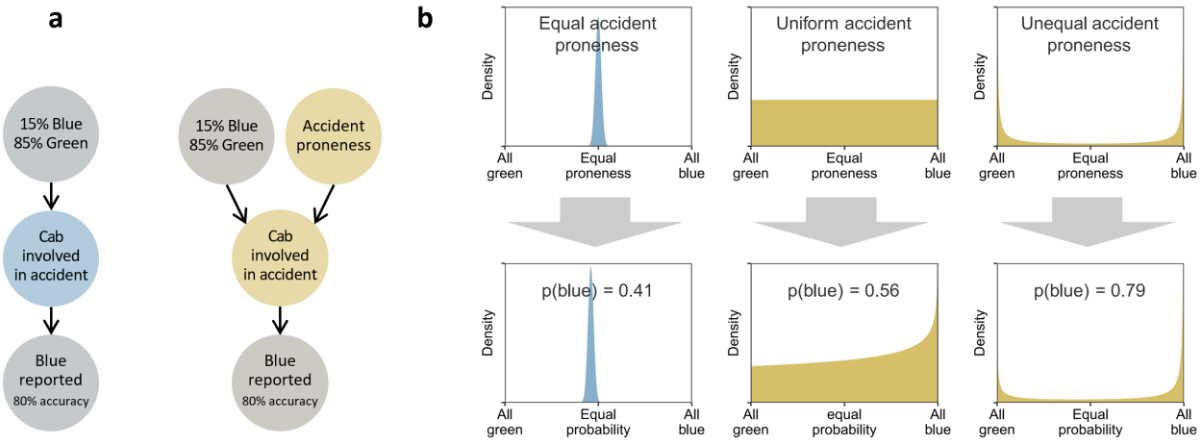


Figure 3. Hierarchical inference neglects incidental but not causal base rates. (**a**) If a causal base rate is given, the non-hierarchical model shown in blue is appropriate. However, given an incidental base rate, accounting for possible differences between the cab companies in proneness to accidents requires the hierarchical model shown in gold. (**b**) Prior assumptions about accident proneness (top row) affect the probability inferred by the hierarchical model that a blue cab was involved in the accident (p(blue); bottom row). 'Accident proneness' denotes the odds that an accident would involve a blue, as opposed to a green cab, had there been equal numbers of blue and green cabs.

accident proneness may or may not differ. Implementing this assumption in the model (**Figure 3b**, middle panel) demonstrates that the probability that the cab was blue should be influenced by the incidental base rate, but not as much as in the case of a causal base rate. In other words, hierarchical inference entails here that the incidental base rate should be partially neglected.

This account of base rate neglect is distinct in key ways from a compelling, alternative account founded on signal detection theory (Birnbaum, 1983). The basic idea of the latter account is that if participants assume the witness is aware of the incidental base rate, then they can already count on the witness taking the base rate into account, and therefore do not need to account for it further. Support for this account comes from a demonstration that base rate neglect indeed decreases once the scenario is modified such that the witness no longer performs a probabilistic judgment (Krynski and Tenenbaum, 2007). In this modified scenario, cab colors fade such that 20% of green cabs appear blue and 20% of blue cabs appear green, whereas the witness reports precisely what they saw and thus has no reason to account for the base rate. However, although this is a compelling explanation for the incidental base rate data, it fails to explain why individuals perform differently when given causal base rates. Our hierarchical-inference explanation predicts this difference because causal base rates account for possible differences in accident proneness between the two cab companies whereas incidental base rates do not.

When are hierarchical inferences rational?

Hierarchical inference is warranted by the assumption of hierarchically organized dependencies between different sets of observations. It is thus irrational whenever a dependency assumed to exist between observations is inconsistent with available evidence. For example, in the case of the halo effect described above, hierarchical inference makes the

assumption that a student's knowledge on different questions is linked together via the student's general level of knowledge on the exam topic. Whereas this particular assumption may be justified, in other cases, a halo effect may result from erroneous assumptions. For instance, a person's physical appearance influences judgment of their personality characteristics (Moore, Filippou, and Perrett, 2011; Wade and DiMaria, 2003), which implies the assumption of a common cause underlying attractiveness and personality. In both cases, hierarchical inferences biases evaluations but only in the latter case the assumption is likely incorrect and the bias irrational.

The suggestion that some biases arise from hierarchical inference that is based on erroneous assumptions raises the question of why people would hold these erroneous assumptions in the first place. Indeed, establishing that people employ a certain inference model requires showing not only that the model produces people's observed behavior, but also explaining why people would hold the assumptions embedded in the model (Geisler et al., 2001; Kemp et al., 2007; Devaine et al., 2014). One way to justify a range of erroneous assumptions is proposed by the theory of ecological rationality (Todd and Gigerenzer, 2012, Hertwig et al., 2021), which posits that organisms rationally develop decision strategies that are suitable to many frequently encountered problems, and by consequence produce behavior that is not suited to infrequently encountered (e.g., experimentally contrived) contexts (Gigerenzer and Brighton, 2009; Gigerenzer, Hertwig and Pachur, 2011). This idea is best exemplified by considering our model of the 'outcome bias'.

An outcome bias is encountered, for instance, when people are asked to evaluate whether a surgeon made the right decision in choosing to perform a surgical procedure (Baron and Hershey, 1988). In this experiment, people are told the expected success rate of the surgery, and so it is irrelevant whether the surgery eventually succeeded or not. Nevertheless, they tend to evaluate the surgeon's decision to operate more highly if the

surgery succeeded. This outcome bias can be interpreted as the product of hierarchical inference assuming that the expected success rate for specific patients deviates from average due to random individual characteristics (e.g., differences in symptoms or in genes; **Figure 2**). Given this assumption, a successful surgery despite a low expected success rate would indicate that for the specific patient the expected success rate was in fact higher than average, and thus, the surgeon's decision to operate was more justified.

The traditional view holds people's evaluations in this experiment to be irrational since people are told that the surgeon did not have access to any additional information beyond the average expected rates of success (Baron and Hershey, 1988). However, in more naturalistic settings, we do not typically have complete knowledge of what information the decision maker has, nor do we know for certain what information can reliably predict a better or worse outcome. Thus, in evaluating real-world medical decisions, we should take into account both known (i.e., expected success rate) and unknown (i.e., variance in success rate between patients) factors. Using outcomes to hierarchically infer the correctness of ours and others' decisions is thus rational and, in fact, essential to how we learn from experience in real life (Hertwig et al., 2021).

Do biases reflect limited or enhanced cognition?

A hierarchical inference account of decision biases sharply diverges from prior accounts in suggesting that biases often result from a more complex, not simpler, form of computation. Almost all literature on this subject has thus far assumed the opposite, namely, that biases arise due to limited cognitive capacity and a need to minimize cognitive costs. This assumption is at the very heart of the idea of bounded rationality (Simon, 1979) and has been a key principle of the heuristics and biases literature (Kahneman, Slovic, and Tversky, 1982). More recently, this idea has been formalized and rationalized as the resource

rationality framework (Dasgupta, Schulz, Goodman, and Gershman, 2018; Dasgupta, Schulz, Tenenbaum, and Gershman, 2020; Gul, Krueger, Callaway, and Griffiths, 2018; Lieder, Griffiths, and Hsu, 2018; Lieder, Griffiths, M Huys, et al., 2018; Polanía, Woodford, and Ruff, 2019). Put simply, this framework views biased decisions as a consequence of the rational deployment of limited cognitive resources to solve decision problems. For instance, to reduce cognitive load and computation time, people may estimate a probability distribution by drawing only a few samples from it (Sanborn and Chater, 2016), or plan only a limited number of steps into the future and beyond this rely on error-prone habits (Keramati, Smittenaar, Dolan, and Dayan, 2016). On this, the resource-rationality framework agrees with the ecological rationality literature, as both view biases as consequences of computationally cheap decision rules (Gigerenzer, Hell and Blank, 1988; Hertwig et al., 2021). By contrast, the hierarchical inference perspective suggests that reaching unbiased decisions can be less costly since a biased decision maker processes contextual variables that an unbiased decision maker can ignore (e.g., a student's typical level of knowledge in the halo effect, and accident proneness in base rate neglect).

If biases indeed arise from a complex and costly form of inference, and not from an attempt to minimize the use of limited resources, then we may expect biases to be diminished in people who invest less effort in solving a decision problem or whose inference capabilities are otherwise compromised. Indeed, recent findings suggest that, rather than being exacerbated, in some cases biases are diminished as a result of reduced neurocognitive function. First, preliminary work indicates that pupillary indices of cognitive load and effort are associated with a greater degree of bias in a range of decision-making tasks (both within and between subjects; Eldar, Felso, Cohen and Niv, 2020). Additionally, anecdotal evidence suggests that lesions to ventromedial prefrontal cortex (vmPFC) are associated with diminished hot-hand illusions and context-dependent biases in value learning (Manohar et al.,

2021). Similarly, a recent study has shown that individuals on the autism spectrum are less susceptible to decision biases because they tend to discount prior context and seemingly irrelevant information when making decisions (Rosenkrantz, D'Mello and Gabrieli, 2021). In all of these cases, the reduction in cognitive biases accompanies changes in behavior that suggest impaired hierarchical inference (e.g., impaired theory of mind in autism; Fields and Glazebrook, 2020; Pezzulo, Rigoli, and Friston, 2018). These findings, along with previous work showing that monetary incentives to act without bias often do not decrease bias (Tversky and Kahneman, 1981b), are consistent with a hierarchical-inference account, and not with the prevailing accounts of biases as means to minimize computational cost.

Of course, the involvement of hierarchical inference in producing a bias would not always predict the bias will intensify with effort, as this prediction depends on algorithmic details that may vary from case to case. Indeed, more broadly, the evidence on the relationship of biases with effort and intact cognitive function is best summarized as mixed (e.g., Alós-Ferrer et al., 2016; Raoelison and De Neys, 2019; Nestler et al., 2008; Igou and Bless, 2007; Diederich et al., 2018; Keramati et al., 2016; Lieder et al., 2018; Epley and Gilovich, 2006; Simmons et al., 2010). This mixed picture coheres with the goal of the present paper, to demonstrate the viability and generativity of a hierarchical inference account to produce explanations of many, but not all, cognitive biases. Ultimately, we believe a complementary set of ideas is needed to comprehensively address the diverse set of cognitive biases, including not only hierarchical inference and resource rationality, but also evolutionary suboptimality and motivated cognition (Williams, 2020). Determining what explanation, or combination of explanations, best suits each instance of a bias requires careful case-by-case study, which we hope the present paper will motivate and inspire. For this purpose, future work could utilize the explicit models outlined here to devise experimental manipulations that would uniquely influence hierarchical inferences.

Concluding remarks

In sum, despite being better informed and more challenging to implement than simple Bayesian inference, hierarchical inference may be responsible for a range of decision-making biases that are often used to highlight the limitations of human reasoning. We propose that the employment of hierarchical inference in these cases is best understood as a way to mitigate uncertainty at the cost of introducing bias. Understanding decision biases through this lens takes the characterization of a bias beyond phenomenological description and reveals the computations and assumptions it reflects. In so doing, the hierarchical inference lens shows how common human biases could arise from fundamental computations that a hierarchically structured brain has evolved to perform (e.g., Friston, 2012; Knill and Pouget, 2004). Studying how these computations are neurally implemented and encoded may thus foster a mechanistic understanding of how biases emerge. Furthermore, the hierarchical inference lens generates novel behavioral predictions concerning people's decisions and their adaptive and maladaptive consequences. It may thus inform both scientists and policy makers with a richer understanding of human decision-making.

Box 3. Generative hierarchical models of decision biases.
We provide here generative hierarchical models justifying each of the biases discussed in the paper. For each model, we describe an example decision query, and probabilistic relationships between variables the decision maker observes (capitalized) and those she needs to infer (marked in bold). Further specified are additional quantities, the estimates of which influence the resulting biases. Priors are left unspecified.

Halo effect

Example: Inferring a student's knowledge on open-ended exam questions from their answers.

Target inference:
$\boldsymbol{\theta}_i \in [0,1]$ – student's knowledge on question *i*

Observed variables:
$A_i \in \{0,1,2,\ldots,n_i\}$ – number of correct features (out of n_i) identified in student's answer to question *i*

Other inferred variables:
$\boldsymbol{\omega} \in [0,1]$ – student's typical knowledge
$\boldsymbol{\kappa} \in \mathbb{R}^+$ – consistency of student's knowledge across questions

Relationships between variables:

$\mathrm{p}(A_i|\boldsymbol{\theta}_i) = \mathrm{Binomial}(A_i|n_i, p = \boldsymbol{\theta}_i)$
$\mathrm{p}(\boldsymbol{\theta}_i|\boldsymbol{\omega},\boldsymbol{\kappa}) = \mathrm{Beta}(\boldsymbol{\theta}_i|\boldsymbol{\omega},\boldsymbol{\kappa})$

Affective evaluation

Example: An animal learns the expected *value* of harvesting fruit from different trees (specific *actions*) in its valley (*general environmental factor*) by harvesting them and observing how much fruit was obtained from each (*observed reward*).

Target inferences:
$\boldsymbol{v}_{i,t} \in \mathbb{R}$ – value of action *i* at time *t*

Observed variables:
$R_{i,t} \in \mathbb{R}$ – observed reward for action *i* at time *t*

Other inferred variables:
$\boldsymbol{g}_t \in \mathbb{R}$ – value of a general environmental factor at time *t*

Relationships between variables:
$\mathrm{p}(R_{i,t}|\boldsymbol{v}_{i,t}) = \mathrm{Normal}(R_{i,t}|\boldsymbol{v}_{i,t}, \sigma_R)$
$\mathrm{p}(\boldsymbol{v}_{i,t}|\boldsymbol{v}_{i,t-1}, \boldsymbol{g}_t, \boldsymbol{g}_{t-1}) = \mathrm{Normal}(\boldsymbol{v}_{i,t}|\boldsymbol{v}_{i,t-1} + (\boldsymbol{g}_t - \boldsymbol{g}_{t-1}), \sigma_v)$
$\mathrm{p}(\boldsymbol{g}_t|\boldsymbol{g}_{t-1}) = \mathrm{Normal}(\boldsymbol{g}_t|\boldsymbol{g}_{t-1}, \sigma_g)$

Additional estimated quantities:
$\sigma_R \in \mathbb{R}^+$ – deviation of individual rewards from expected value
$\sigma_v \in \mathbb{R}^+$ – independent volatility, of specific action values
$\sigma_g \in \mathbb{R}^+$ – common volatility, of general environmental factor

Attribute framing

Example: Estimating the quality of a computer (*product value*) after a friend (*experimenter*) tells you from her experience the proportion of times the computer did (*positively framed attribute*) or did not (*negatively framed attribute*) handled tasks efficiently.

Target inference:
$\boldsymbol{v} \in \mathbb{R}$ – a product's value

Observed variables:
$A \in \mathbb{R}$ – an attribute of the product
$F \in \{0 = \mathrm{negative}, 1 = \mathrm{positive}\}$ – frame

Other inferred variables:
$\boldsymbol{k} \in \mathbb{R}$ – experimenter's evaluation of the product

Relationships between variables:
$\mathrm{p}(F = 1|\boldsymbol{k}) = \mathrm{logistic}(\beta \boldsymbol{k})$
$\mathrm{p}(\boldsymbol{k}|\boldsymbol{v}) = \mathrm{Normal}(\boldsymbol{k}|\boldsymbol{v}, \sigma_k)$
$\mathrm{p}(A|\boldsymbol{v}) = \mathrm{Normal}(A|\boldsymbol{v}, \sigma_A)$

Additional estimated quantities:
$\beta \in \mathbb{R}^+$ – influence of experimenter's evaluation on frame
$\sigma_k \in \mathbb{R}^+$ – reliability of experimenter's evaluation
$\sigma_A \in \mathbb{R}^+$ – relation between value and attribute

Anchoring bias

Example: Estimating the GDP (*quantity of interest*) of the US using both common knowledge (*known facts)* and a related value (*anchor*) provided by an experimenter.

Target inference:
$\boldsymbol{q} \in \mathbb{R}$ – a quantity of interest

Observed variables:
$A \in \mathbb{R}$ – anchor
$\vec{F} \in \mathbb{R}^n$ – known facts of size *n*

Other inferred variables:
$\boldsymbol{k} \in \mathbb{R}$ – experimenter's estimate of the quantity

Relationships between variables:
$\mathrm{p}(A|\boldsymbol{k}) = \mathrm{Normal}(A|\boldsymbol{k}, \sigma_A)$
$\mathrm{p}(\boldsymbol{k}|\boldsymbol{q}) = \mathrm{Normal}(\boldsymbol{k}|\boldsymbol{q}, \sigma_k)$
$\mathrm{p}(\vec{F}|\boldsymbol{q}) = \mathrm{Normal}(\vec{F}|\boldsymbol{q}, \Sigma_F)$

Additional estimated quantities:
$\sigma_k \in \mathbb{R}^+$ – reliability of experimenter's estimate
$\sigma_A \in \mathbb{R}^+$ – deviation of anchor from experimenter's knowledge
$\Sigma_F \in \mathbb{R}^{n \times n}$ – similarity between known facts

Hot hand illusion

Example: Inferring a basketball player's changing *ability* to make baskets based on their recent history of made and missed baskets.

Target inference:
$\boldsymbol{a}_{i,t} \in \mathbb{R}$ – ability of basketball player *i* at time *t*

Observed variables:
$M_{i,t} \in \{0 = \mathrm{miss}, 1 = \mathrm{make}\}$ – make or miss by basketball player *i* at time *t*

Relationships between variables:
$\mathrm{p}(M_{i,t} = 1|\boldsymbol{a}_{i,t}) = \mathrm{logistic}(\boldsymbol{a}_{i,t})$
$\mathrm{p}(\boldsymbol{a}_{i,t}|\boldsymbol{a}_{i,t-1}) = \mathrm{Normal}(\boldsymbol{a}_{i,t}|\boldsymbol{a}_{i,t-1}, \sigma_a)$

Additional estimated quantities:
$\sigma_a \in \mathbb{R}^+$ – volatility of basketball players' ability

Base-rate Neglect

Example: Judging the probability that a car involved in an accident belonged to the blue cab company, as opposed to the green cab company, given that only 15% of cabs are blue (*incidental base rate*) and that a moderately-reliable (i.e., 80% accurate) eyewitness said they saw a blue cab.

Target inference:
$\boldsymbol{c} \in \{0 = \text{green}, 1 = \text{blue}\}$ – cab involved in accident

Observed variables:
$R \in \{0 = \text{green}, 1 = \text{blue}\}$ – cab reported by eyewitness
$I_{\text{blue}} \in [0,1]$ – incidental base rate of blue cab
$I_{\text{green}} = 1 - I_{\text{blue}}$ – incidental base rate of green cab

Other inferred variables:
$\boldsymbol{P}_{\text{blue}} \in [0,1]$ – relative accident proneness of blue cab
$\boldsymbol{P}_{\text{green}} = 1 - \boldsymbol{P}_{\text{blue}}$ – relative accident proneness of green cab

Relationships between variables:
$\text{p}(\boldsymbol{c} = 1 | I_{\text{blue}}, I_{\text{green}}) = \frac{I_{\text{blue}}\boldsymbol{P}_{\text{blue}}}{I_{\text{blue}}\boldsymbol{P}_{\text{blue}} + I_{\text{green}}\boldsymbol{P}_{\text{green}}}$
$\text{p}(R = 1 | \boldsymbol{c}) = 0.8\boldsymbol{c} + 0.2\,(1 - \boldsymbol{c})$

Outcome bias

Example: Determining whether a surgeon made the correct decision to perform surgery (*decision correctness*) based on the known success rate of the surgery (*fixed expected outcome*), what might be known about the individual patient's characteristics (*random effects*), and the outcome of the surgery.

Target inference:
$\boldsymbol{d} \in \{0 = \text{incorrect}, 1 = \text{correct}\}$ – decision correctness

Observed variables:
$O \in \{0 = \text{unsuccessful}, 1 = \text{successful}\}$ – outcome
$F \in [0,1]$ – fixed expected outcome

Other inferred variables:
$\boldsymbol{r} \in \mathbb{R}$ – random effects

Relationships between variables:
$\text{p}(O = 1 | F, \boldsymbol{r}) = \text{logistic}(\text{logit}(F) + \boldsymbol{r})$
$\boldsymbol{d} = \begin{cases} 1 & \text{if p}(O = 1 | F, \boldsymbol{r}) > \theta \\ 0 & \text{else} \end{cases}$

Additional estimated quantities:
$\theta \in [0,1]$ – threshold for evaluating a decision as correct

Availability heuristic

Example: Guessing the chance a massive flood will occur somewhere in North America, or conversely, a massive flood will occur due to an earthquake in California.

Target inference:
$\boldsymbol{f}_i \in \mathbb{R}^+$ – frequency of flood of type *i*

Observed variables:
$R_i \in \mathbb{R}^+$ – rate of information coming to mind about flood of type *i*

Other inferred variables:
$\boldsymbol{n}_i \in \mathbb{Z}^+$ – number of previous encounters with flood of type *i*

Relationships between variables:

$\text{p}(R_i | \boldsymbol{n}_i) = \text{Gamma}(R_i | \mu = \beta \boldsymbol{n}_i, \sigma)$
$\text{p}(\boldsymbol{n}_i | \boldsymbol{f}_i) = \text{Poisson}(\boldsymbol{n}_i | \boldsymbol{f}_i + d)$

Additional estimated quantities:
$\beta \in \mathbb{R}^+$ – effect of number of encounters on average rate of information coming to mind
$\sigma \in \mathbb{R}^+$ – variability in rate of information due to factors other than number of encounters
$d \in \mathbb{R}^+$ – variability in rate of encounters due to factors other than the flood's frequency

References

Alós-Ferrer, C., Garagnani, M., & Hügelschäfer, S. (2016). Cognitive reflection, decision biases, and response times. Frontiers in psychology, 7, 1402.

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. Acta Psychologica, 44(3), 211–233.

Baron, J., and Hershey, J. C. (1988). Outcome bias in decision evaluation. Journal of Personality and Social Psychology, 54(4), 569–579. https://doi.org/10.1037/0022-3514.54.4.569

Bartolo, R., and Averbeck, B. B. (2021). Inference as a fundamental process in behavior. *Current Opinion in Behavioral Sciences*, 38, 8-13.

Behrens, T. E., Hunt, L. T., Woolrich, M. W., and Rushworth, M. F. (2008). Associative learning of social value. *Nature*, 456(7219), 245-249.

Benrimoh, D., Parr, T., Vincent, P., Adams, R. A., and Friston, K. (2018). Active inference and auditory hallucinations. *Computational Psychiatry (Cambridge, Mass.)*, *2*, 183–204. https://doi.org/10.1162/cpsy_a_00022

Birnbaum, M. H. (1983). Base rates in Bayesian inference: Signal detection analysis of the cab problem. *The American Journal of Psychology*, 85-94.

Dasgupta, I., Schulz, E., Goodman, N. D., and Gershman, S. J. (2018). Remembrance of inferences past: Amortization in human hypothesis generation. *Cognition*, *178*, 67–81.

Dasgupta, I., Schulz, E., Tenenbaum, J. B., and Gershman, S. J. (2020). A theory of learning to infer. *Psychological Review*, *127*(3), 412–441. https://doi.org/10.1037/rev0000178

De Berker, A. O., Rutledge, R. B., Mathys, C., Marshall, L., Cross, G. F., Dolan, R. J., and Bestmann, S. (2016). Computations of uncertainty mediate acute stress responses in humans. *Nature communications*, 7(1), 1-11.

De Lange, F. P., Heilbron, M., and Kok, P. (2018). How do expectations shape perception?. *Trends in Cognitive Sciences*, 22(9), 764-779

Devaine, M., Hollard, G., and Daunizeau, J. (2014). Theory of mind: did evolution fool us? *PloS One*, *9*(2), e87619.

Diaconescu, A. O., Stecy, M., Kasper, L., Burke, C. J., Nagy, Z., Mathys, C., and Tobler, P. N. (2020). Neural arbitration between social and individual learning systems. *ELife*, *9*. https://doi.org/10.7554/eLife.54051

Diederich, A., Wyszynski, M., and Ritov, I. (2018). Moderators of framing effects in variations of the Asian Disease problem: Time constraint, need and disease type. *Judgment and decision making*, *13*(6), 529.

Eldar, E., Felso, V., Cohen, J. D., and Niv, Y. (2021). A pupillary index of susceptibility to decision biases. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-020-01006-3

Eldar, E., Rutledge, R. B., Dolan, R. J., and Niv, Y. (2016). Mood as representation of momentum. *Trends in Cognitive Sciences*, *20*(1), 15–24.

Epley, N., and Gilovich, T. (2006). The anchoring-and-adjustment heuristic: Why the adjustments are insufficient. Psychological science, 17(4), 311-318.

Fields, C., and Glazebrook, J. F. (2020). Information flow in context-dependent hierarchical Bayesian inference. *Journal of Experimental & Theoretical Artificial Intelligence*, 1–32.

Fradkin, I., Adams, R. A., Parr, T., Roiser, J. P., and Huppert, J. D. (2020). Searching for an anchor in an unpredictable world: A computational model of obsessive compulsive disorder. *Psychological Review*, *127*(5), 672–699. https://doi.org/10.1037/rev0000188

Fradkin, I., Ludwig, C., Eldar, E., and Huppert, J. D. (2020). Doubting what you already know: Uncertainty regarding state transitions is associated with obsessive compulsive symptoms. *PLoS Computational Biology*, *16*(2), e1007634.

Friston, K. (2012). The history of the future of the Bayesian brain. *Neuroimage, 62*(2), 1230-1233.

Fudenberg, D., Levine, D. K., & Maniadis, Z. (2012). On the robustness of anchoring effects in WTP and WTA experiments. *American Economic Journal: Microeconomics, 4*(2), 131-45.

Geisler, W. S., Perry, J. S., Super, B. J., & Gallogly, D. P. (2001). Edge co-occurrence in natural images predicts contour grouping performance. *Vision research, 41*(6), 711-724.

Gigerenzer, G. (2018). The bias bias in behavioral economics. *Review of Behavioral Economics, 5*(3-4), 303-336.

Gigerenzer, G., and Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in cognitive science, 1*(1), 107-143.

Gigerenzer, G., Hell, W., and Blank, H. (1988). Presentation and content: The use of base rates as a continuous variable. *Journal of Experimental Psychology: Human Perception and Performance*, 14(3), 513.

Gigerenzer, G. E., Hertwig, R. E., and Pachur, T. E. (2011). *Heuristics: The foundations of adaptive behavior*. Oxford University Press.

Gilovich, T., Griffin, D., and Kahneman, D. (2002). *Heuristics and biases: the psychology of intuitive judgment* (T. Gilovich, D. Griffin, and D. Kahneman, Eds.). Cambridge University Press. https://doi.org/10.1017/CBO9780511808098

Gilovich, T., Vallone, R., and Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology*, *17*(3), 295–314.

Glaze, C. M., Filipowicz, A. L. S., Kable, J. W., Balasubramanian, V., and Gold, J. I. (2018). A bias–variance trade-off governs individual differences in on-line learning in an

unpredictable environment. *Nature Human Behaviour*, *2*(3), 213–224. https://doi.org/10.1038/s41562-018-0297-4

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., and Tenenbaum, J. B. (2010). Probabilistic models of cognition: exploring representations and inductive biases. *Trends in Cognitive Sciences*, *14*(8), 357–364. https://doi.org/10.1016/j.tics.2010.05.004

Grüne-Yanoff, T. (2017). Reflections on the 2017 Nobel memorial prize awarded to Richard Thaler. *Erasmus Journal for Philosophy and Economics.*

Gul, S., Krueger, P. M., Callaway, F., and Griffiths, T. L. (2018). Discovering rational heuristics for risky choice. *KogWis.*

Guomei, Z., and Qicheng, J. (2003). Psychologist Daniel Kahneman Wins 2002 Nobel Prize in Economics. *Advances in Psychological Science.*

Gweon, H. (2021). Inferential social learning: Cognitive foundations of human social learning and teaching. *Trends in Cognitive Sciences,* 25(10), 896-910.

Hertwig R, Leuker C, Pachur T, Spiliopoulos L, Pleskac TJ. (2021). Studies in ecological rationality. *Topics in Cognitive Science*. DOI:10.1111/tops.12567.

Hesp, C., Smith, R., Parr, T., Allen, M., Friston, K. J., and Ramstead, M. J. D. (2021). Deeply felt affect: the emergence of valence in deep active inference. *Neural Computation*, *33*(2), 398–446.

Hintzman, D. L. (1976). Repetition and memory. *Psychology of learning and motivation*, *10*, 47-91.

Igou, E. R., & Bless, H. (2007). On undesirable consequences of thinking: Framing effects as a function of substantive processing. *Journal of Behavioral Decision Making*, *20*(2), 125-142.

Ioannidis, K., Offerman, T., & Sloof, R. (2020). On the effect of anchoring on valuations when the anchor is transparently uninformative. *Journal of the Economic Science Association*, *6*(1), 77-94.

Jacowitz, K. E., and Kahneman, D. (1995). Measures of anchoring in estimation tasks. *Personality and Social Psychology Bulletin*, *21*(11), 1161-1166.

Kahneman, D., Slovic, S. P., Slovic, P., and Tversky, A. (Eds.). (1982). Judgment under uncertainty: Heuristics and biases. Cambridge university press.

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York: Penguin/Robinson.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. Heuristics and biases: The psychology of intuitive judgment, 49, 81.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental science, 10*(3), 307-321.

Keramati, M., Smittenaar, P., Dolan, R. J., and Dayan, P. (2016). Adaptive integration of habits into depth-limited planning defines a habitual-goal-directed spectrum. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(45), 12868–12873.

Knill, D. C., and Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences*, *27*(12), 712–719.

Krynski, T. R., and Tenenbaum, J. B. (2007). The role of causality in judgment under uncertainty. *Journal of Experimental Psychology: General*, 136(3), 430.

Lawson, R. P., Mathys, C., and Rees, G. (2017). Adults with autism overestimate the volatility of the sensory environment. *Nature Neuroscience*, *20*(9), 1293–1299.

Lee, M. D., and Newell, B. R. (2011). Using hierarchical Bayesian methods to examine the tools of decision-making. *Judgement and Decision Making.*

Levin, I. P., Schneider, S. L., & Gaeth, G. J. (1998). All frames are not created equal: A typology and critical analysis of framing effects. *Organizational behavior and human decision processes*, *76*(2), 149-188.

Lieder, F., Griffiths, T. L., and Hsu, M. (2018). Overrepresentation of extreme events in decision making reflects rational use of cognitive resources. *Psychological Review*, *125*(1), 1–32.

Lieder, F., Griffiths, T. L., M Huys, Q. J., and Goodman, N. D. (2018). Empirical evidence for resource-rational anchoring and adjustment. *Psychonomic Bulletin & Review*, *25*(2), 775–784.

Mandel, D. R. (2014). Do framing effects reveal irrational choice? Journal of Experimental Psychology: General, 143(3), 1185.

Manohar, S., Lockwood, P., Drew, D., Fallon, S. J., Chong, T. T.-J., Jeyaretna, D. S., … Husain, M. (2021). Reduced decision bias and more rational decision making following ventromedial prefrontal cortex damage. *Cortex*, *138*, 24–37.

Michely, J., Eldar, E., Martin, I. M., and Dolan, R. J. (2020). A mechanistic account of serotonin's impact on mood. *Nature Communications*, 11(1), 2335. https://doi.org/10.1038/s41467-020-16090-2

Miller, J. B., and Sanjurjo, A. (2018). Surprised by the hot hand fallacy? A truth in the law of small numbers. *Econometrica : Journal of the Econometric Society*, *86*(6), 2019–2047. https://doi.org/10.3982/ECTA14943

Moore, F. R., Filippou, D., and Perrett, D. I. (2011). Intelligence and attractiveness in the face: Beyond the attractiveness halo effect. *Journal of Evolutionary Psychology*, *9*(3), 205–217. https://doi.org/10.1556/JEP.9.2011.3.2

Nestler, S., Blank, H., & von Collani, G. (2008). Hindsight bias doesn't always come easy: Causal models, cognitive effort, and creeping determinism. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(5), 1043.

Pezzulo, G., Rigoli, F., and Friston, K. J. (2018). Hierarchical active inference: A theory of motivated control. *Trends in Cognitive Sciences*, *22*(4), 294–306. https://doi.org/10.1016/j.tics.2018.01.009

Piray, P., and Daw, N. D. (2020). A simple model for learning in volatile environments. *PLoS Computational Biology*, *16*(7), e1007963. https://doi.org/10.1371/journal.pcbi.1007963

Polanía, R., Woodford, M., and Ruff, C. C. (2019). Efficient coding of subjective value. *Nature Neuroscience*, *22*(1), 134–142. https://doi.org/10.1038/s41593-018-0292-0

Powers, A. R., Mathys, C., and Corlett, P. R. (2017). Pavlovian conditioning-induced hallucinations result from overweighting of perceptual priors. *Science*, כ*357*(6351), 596–600. https://doi.org/10.1126/science.aan3458

Qiu, C., Luu, L., and Stocker, A. A. (2020). Benefits of commitment in hierarchical inference. *Psychological Review*, *127*(4), 622–639. https://doi.org/10.1037/rev0000193

Raoelison, M., & De Neys, W. (2019). Do we de-bias ourselves?: The impact of repeated presentation on the bat-and-ball problem. *Judgment and Decision making*, *14*(2), 170.

Reed, E. J., Uddenberg, S., Suthaharan, P., Mathys, C. D., Taylor, J. R., Groman, S. M., and Corlett, P. R. (2020). Paranoia as a deficit in non-social belief updating. *ELife*, *9*. https://doi.org/10.7554/eLife.56345

Ritzwoller, D. M., & Romano, J. P. (2022). Uncertainty in the hot hand fallacy: Detecting streaky alternatives to random bernoulli sequences. The Review of Economic Studies, 89(2), 976-1007.

Rozenkrantz, L., D'Mello, A. M., and Gabrieli, J. D. E. Enhanced rationality in autism spectrum disorder. *Trends in Cognitive Sciences.* 2021 Aug;25(8):685–96.

Sanborn, A. N., and Chater, N. (2016). Bayesian brains without probabilities. *Trends in cognitive sciences, 20*(12), 883-893.

Schneider, B., and Koenigs, M. (2017). Human lesion studies of ventromedial prefrontal cortex. *Neuropsychologia, 107*, 84–93.

Schustek, P., Hyafil, A., and Moreno-Bote, R. (2019). Human confidence judgments reflect reliability-based hierarchical integration of contextual information. *Nature Communications*, *10*(1), 5430. https://doi.org/10.1038/s41467-019-13472-z

Schwarz, N., & Clore, G. L. (1983). Mood, misattribution, and judgments of well-being: informative and directive functions of affective states. *Journal of personality and social psychology*, *45*(3), 513.

Sher, S. & McKenzie, CRM. 2006. "Information leakage from logically equivalent frames". Cognition. 101: 467–494.

Siegel, J. Z., Mathys, C., Rutledge, R. B., and Crockett, M. J. (2018). Beliefs about bad people are volatile. *Nature Human Behaviour*, *2*(10), 750–756.

Simmons, J. P., LeBoeuf, R. A., and Nelson, L. D. (2010). The effect of accuracy motivation on anchoring and adjustment: Do people adjust from provided anchors? Journal of personality and social psychology, 99(6), 917.

Simon, H. A. (1979). Rational decision making in business organizations. *The American Economic Review*, 69(4), 493-513.

Slovic, P., Finucane, M. L., Peters, E., and MacGregor, D. G. (2007). The affect heuristic. *European Journal of Operational Research*, *177*(3), 1333–1352. https://doi.org/10.1016/j.ejor.2005.04.006

Smith, R., Thayer, J. F., Khalsa, S. S., and Lane, R. D. (2017). The hierarchical basis of neurovisceral integration. *Neuroscience and Biobehavioral Reviews*, *75*, 274–296. https://doi.org/10.1016/j.neubiorev.2017.02.003

Summerfield, C., and Tsetsos, K. (2012). Building Bridges between Perceptual and Economic Decision-Making: Neural and Computational Mechanisms. *Frontiers in Neuroscience, 6*, 70. https://doi.org/10.3389/fnins.2012.00070

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., and Goodman, N. D. (2011). How to grow a mind: statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285. https://doi.org/10.1126/science.1192788

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of applied psychology*, *4*(1), 25.

Todd, P. M., and Gigerenzer, G. E. (2012). Ecological rationality: Intelligence in the world. Oxford University Press.

Tversky, A., and Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science, 185*(4157), 1124-1131.

Tversky, A., and Kahneman, D. (1980). Causal schemas in judgments under uncertainty. In Fishbein, M. (Ed.), *Progress in social psychology* (Vol. 1, pp. 49–72). Hillsdale, NJ: Lawrence Erlbaum.

Tversky, A., and Kahneman, D. (1981a). *Evidential impact of base rates*. Stanford University.

Tversky, A., and Kahneman, D. (1981b). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453–458.

Tversky, A., and Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological review, 90*(4), 293.

van Ravenzwaaij, D., Moore, C. P., Lee, M. D., and Newell, B. R. (2014). A hierarchical Bayesian modeling approach to searching and stopping in multi-attribute judgment. *Cognitive Science, 38*(7), 1384–1405. https://doi.org/10.1111/cogs.12119

Wade, T. J., and DiMaria, C. (2003). Weight halo effects: Individual differences in perceived life success as a function of women's race and weight. Sex Roles.

Williams, D. (2020). Epistemic Irrationality in the Bayesian Brain. *The British Journal for the Philosophy of Science*.

Wilson, T. D., Houston, C. E., Etling, K. M., and Brekke, N. (1996). A new look at anchoring effects: basic anchoring and its antecedents. Journal of Experimental Psychology: General, 125(4), 387.

Declarations

Funding (information that explains whether and by whom the research was supported)

Paul B. Sharp is supported by postdoctoral fellowship from the Fulbright Association (PS00318453)

Eran Eldar is supported by NIH grants R01MH124092 and R01MH125564, ISF grant 1094/20 and US 1336 Israel BSF grant 2019801.

Conflicts of interest/Competing interests (include appropriate disclosures): No conflicts.

Ethics approval (include appropriate approvals or waivers): Not applicable.

Consent to participate (include appropriate statements): Not applicable.

Consent for publication (include appropriate statements): Not applicable.

Availability of data and materials (data transparency): Not applicable.

Code availability (software application or custom code): Not applicable.