PERSPECTIVE
https://doi.org/10.1038/s42256-019-0038-z
nature machine intelligence

Lessons for artificial intelligence from the study of natural stupidity

Alexander S. Rich[1,2*] and Todd M. Gureckis[1]

Artificial intelligence and machine learning systems are increasingly replacing human decision makers in commercial, healthcare, educational and government contexts. But rather than eliminate human errors and biases, these algorithms have in some cases been found to reproduce or amplify them. We argue that to better understand how and why these biases develop, and when they can be prevented, machine learning researchers should look to the decades-long literature on biases in human learning and decision-making. We examine three broad causes of bias—small and incomplete datasets, learning from the results of your decisions, and biased inference and evaluation processes. For each, findings from the psychology literature are introduced along with connections to the machine learning literature. We argue that rather than viewing machine systems as being universal improvements over human decision makers, policymakers and the public should acknowledge that these system share many of the same limitations that frequently inhibit human judgement, for many of the same reasons.

Amos Tversky said, "My colleagues, they study artificial intelligence; me, I study natural stupidity." Artificial intelligence (AI) and machine learning systems are quickly reaching or surpassing human performance in many domains, from playing complex games[1] to recognizing objects in natural scenes[2]. Because of this, AI systems seem poised to take an ever increasing role in commercial, healthcare, educational and governmental decisions. However, as AI enters into more sensitive and societally important areas, there is growing evidence that automated systems do not always unambiguously improve on human decision-making. In some cases, machine learning models are simply inaccurate, such as the discontinued Google Flu Trends[3]. But in many, worrying cases, the outputs of a statistically biased model create socially biased outcomes, perpetuating and amplifying the very problems that machine learning seemingly promises to prevent[4,5].

These biases can develop over a remarkably wide range of domains and, in an ironic twist, echo the supoptimalities in decision-making, learning and reasoning that have been documented in humans (Table 1). For instance, application-screening algorithms have been shown to directly encode the biases of past human reviewers and perpetuate discrimination against women and minority groups[6]. A recently developed commercial facial recognition software package made more mistakes identifying the gender of darker-skinned than lighter-skinned individuals, possibly because most photographic training sets under-represent darker-skinned people and women[7]. Predictive policing algorithms that send officers to where crime has occurred in the past may cause feedback loops of over-policing[8–10], in a manner similar to those that enforce stereotypes in individuals[11]. Even an algorithm designed to detect potholes using smartphone data can become biased away from finding them in neighbourhoods with low-income and older residents[12], because algorithms, like humans, can undercount rare events when data is sparse[13].

The dialogue between research on machine intelligence and human psychology and neuroscience can be traced to the origins of the field of AI[14–17] and continues today[18,19]. Historically, AI researchers have looked to the successes of human cognition when seeking improvements to their algorithms. But to address the growing issues of bias in AI, it may be equally valuable to look at the pitfalls and failures of cognition—what Amos Tversky playfully referred to as 'natural stupidity'.

Several strands of psychology have studied the ways in which people's choices and inferences are biased, suboptimal or irrational. This includes Nobel-prize-winning work in the psychology of judgement and decision-making around the concept of 'heuristics and biases'[20]. Beyond simply documenting these cases, psychologists have produced numerous insights into the conditions under which these biases manifest, their sources, and when they are avoidable, inevitable or even helpful. Of particular relevance to machine learning is psychological research that analyses human biases from the perspective of rationality and computational constraints (although this is just one of a range of views about human rationality). This research demonstrates how biases, while still harmful, can be understood in terms of the structure of the environment and the agent's goals. It is likely that lessons drawn from this literature can offer useful guide-posts for developing and improving machine learning algorithms. The goal of this Perspective is to establish links from research on human biases to the emerging field of algorithmic bias. In particular, we highlight several well-known examples where understanding the pitfalls in human psychology may help machine learning systems avoid similar fates.

The danger of small data

Unlike machine learning algorithms, which may be trained on millions of examples, humans have to draw inferences from their limited personal experience. In many ways, people are remarkably adept at this task. Machine learning systems are only beginning to approach human performance on learning concepts from a single training episode[21], and major debates in linguistics have centred on understanding how children learn the rules of their language given how few grammatical utterances they hear[22].

Along with exploring these successes, psychologists have also documented how small data, or small samples, frequently warp human judgement. One example is research on how people estimate the frequency of rare events. With small samples, rare events may often be unobserved even if on average the samples are statistically

[1]Department of Psychology, New York University, New York, NY, USA. [2]Flatiron Health, New York, NY, USA. *e-mail: asr443@nyu.edu

Table 1 | Learning and decision-making biases observed in humans, and their application to machine learning

Human bias	Machine learning application
Under-weighting rare events and illusory correlations	Practitioners should check for underestimation of rare events, and whether individuals or groups with little data are treated differently from those with more. These biases can be corrected by collecting more data about rare groups, using different decision policies for rare groups, using less extreme priors or less regularization, or falling back to simple heuristics or non-machine learning solutions when data is sparse.
Hot stove effect and attentional learning trap	If feedback is choice-dependent, the hot stove effect may occur. Check for bias against options with high variance, and bias towards options that are novel or where feedback is not choice-dependent. To reduce bias (at the cost of short-term performance), use random actions or other reinforcement learning algorithms to increase exploration, or train the model only on data where feedback was not choice-dependent. Diagnosing the attentional learning trap and other extensions of the hot stove effect in machine learning algorithms is difficult, but the solutions will resemble those for the hot stove effect, and may also include modifying the model to reduce rapid generalization.
Reference-dependent risk preferences	Models may prefer more variable options when the set of options produces lower-than-average outcomes, and more stable options when the options produce higher-than-average outcomes. Model specification can be adjusted to increase or decrease this effect.
Tallying and other fast and frugal heuristics	Tallying and similar decision rules appear biased, but can sometimes work nearly as well as advanced machine learning algorithms. These approaches can play a role in expanding 'interpretable machine learning'.

unbiased. In lab experiments, this phenomenon causes people to seemingly under-weight low-probability outcomes in gambles where they have seen only a few outcomes[13], despite over-weighting rare events when given full information about the outcome distribution[23]. In daily life, people are often overly optimistic about how unlikely they are to experience rare negative events such as contracting a given disease, possibly because their sample of personal acquaintances may contain no instances of the event[24]. There is some evidence that people are likewise pessimistic about rare positive events[25].

The under-weighting of rare events extends beyond one-off decisions and estimation to repeated-choice settings in which people make a series of consequential decisions. In these scenarios, the effect of small samples can lead people to prefer a risky option over a safe one because the negative, potentially catastrophic, outcome of the risky choice has not been observed. This pattern is seen both in controlled experimental settings[26] and in high-stakes real-world settings such as driving a car[27]. Intriguingly, there is recent evidence that people under-weight rare negative events even when such events have been observed, because they infer false temporal patterns in when the event will occur and thus overestimate their safety[28].

The difficulty of learning from small data extends not just to estimating the frequency of rare events, but also to estimating the properties of rare groups or individuals. For example, in a classic social psychology experiment, participants are brought into a laboratory and observe members of a majority and a minority group performing positive or negative actions[29]. Most individuals observed are from the majority group, and most observed actions are positive, but there is no actual correlation between group membership and the action performed. Despite this, participants generally report perceiving an association between majority group membership and positive actions, an example of a phenomenon referred to as illusory correlation. Illusory correlation has been proposed as a factor that leads to negative stereotypes of minority and under-represented groups[29,30].

While seemingly a uniquely human bias, illusory correlations can also be explained as a product of several generic aspects of learning and information processing[31,32]. One potential cause, of particular interest to the machine learning community, is Bayesian belief updating. In the experiment described above, for example, if participants begin with a low prior belief of the proportion of positive actions in a group, then their beliefs in each group will become more positive over time, but will become positive more quickly for the frequently experienced majority group[33]. In general, as shown in Fig. 1, the believed mean value of a variable for a smaller group will remain closer to prior beliefs than that of a larger group, even

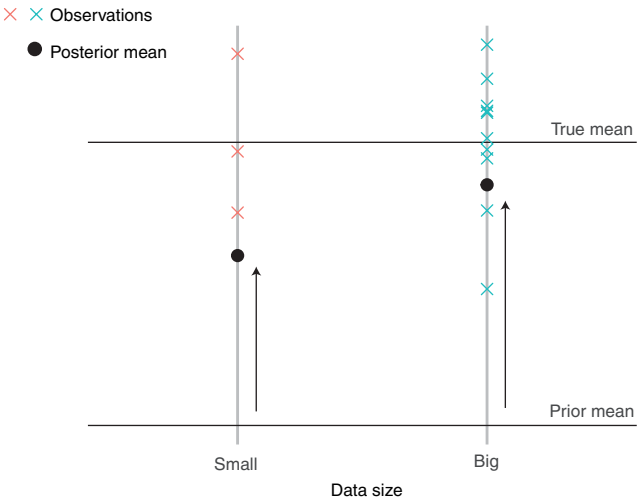

Fig. 1 | In illusory correlations, an agent mistakenly comes to believe that there is a correlation between a variable of interest and membership to a larger group (or more data-rich group or individual). This occurs because more data can shift the agent's posterior beliefs farther from its prior beliefs (arrows). In this figure, the posterior mean for the 'big' data group is higher than that for the 'small' data group, even though data from both groups were generated as noisy observations from the same true mean.

if the two groups have the same true value. This mechanism means that illusory correlations can emerge not just when a group or individual is rare, but also when there are reasons for a group to be undercounted in a dataset. For example, if a classroom has an even number of boys and girls, but the girls answer fewer questions, the teacher may come to believe that the boys have higher ability simply owing to the small sample of questions answered by girls[34].

The machine learning community has often overcome the problems of small samples by assembling sufficiently large datasets[35]. But there will always be some situations, and some individuals, about which data is sparse, and in these situations it is important to understand the biases that small data can induce[36]. Bayesian estimation and other forms of regularization and shrinkage common in machine learning[37] might easily exhibit an illusory correlation in which individuals with little data are treated differently from individuals with a lot. For example, if a social service is algorithmically targeted at the 5% of individuals with the greatest estimated need, individuals about whom little data has been collected may never

have extreme enough estimates to receive the service, regardless of their true degree of need. Similarly, rare events can be missed in under-represented individuals or communities. When Boston released its Street Bump smartphone app to detect potholes using the GPS and accelerometer data of drivers, the city faced the challenge of ensuring detection in neighbourhoods with low-income and older residents, who are far less likely to own smartphones[12].

There are no easy solutions for the biases introduced by small samples, but several approaches to reducing them exist, each with their own costs (see Table 1). Adjusting predictions on the basis of an individual's group membership may be effective in some cases, but can risk replacing the subtle discrimination of an ostensibly neutral algorithm with direct discrimination on a sensitive attribute[6,38,39]. Using less extreme model priors will also reduce illusory correlation, but at the cost of reducing model performance over all. Recent work shows that collecting more data about rare groups can reduce algorithmic discrimination[40], but data collection can be expensive or infeasible. Finally, it may make sense to fall back to simpler heuristic models[41], or non-machine learning approaches, in cases where data for a more advanced model is insufficient. For example, the city of Boston could consider sending out municipal workers to survey the road condition in areas where smartphone penetration is low, while using machine learning in areas where it is high. What the long history of psychological research makes clear is that small sample biases are hard to fix, and that they occur in a wide range of contexts and types of decision. This means that awareness and consideration of them is crucial when designing machine learning systems in situations where some individuals and groups might produce more data than others[36].

Learning from what you choose

Much of human experience can be thought of not as purely observational, but as choice-contingent, with people learning the consequences of only the actions they actually take. For example, we learn about only the meals we have eaten rather than those we pass up. Machine learning systems also often take actions in the world and receive data that is contingent on the consequences of those actions, in domains including robotics, advertising, insurance, finance and operations. Unfortunately, learning from the actions one chooses, like learning from small samples, can produce biases and false beliefs.

The prototypical bias caused by the choice-contingent feedback is the 'hot stove effect', first described by organizational theorists Denrell and March[42]. The hot stove effect is the tendency for a decision maker to develop overly negative estimates of an action with variable outcomes and come to avoid it, even if its average outcome is actually positive. This occurs because of an asymmetry in the consequences of incorrect-positive and incorrect-negative beliefs, as shown in Fig. 2. If the initial feedback suggests that the action has positive consequences, then the decision maker will continue choosing it and will find out whether the early learning was wrong (Fig. 2a). But if the initial observations suggest that the action is negative, then the decision maker will begin to avoid that action and thus never receive corrective feedback. The agent's beliefs can diverge indefinitely from what they would be if the agent continued choosing the option and received 'complete' information (Fig. 2b). Thus, for example, because people have the (partial) ability to avoid other people whom they do not like, negative first impressions tend to last longer than positive first impressions[11].

The hot stove effect, although seemingly very simple, can explain a wide range of behaviours and biases. It can produce apparent risk aversion, because high-variance actions are more likely to produce the effect than are low-variance ones[42]. It can contribute to negative beliefs about social outgroups because interacting with and learning about ingroup members, such as family members, is generally unavoidable, whereas learning about outgroup members is often choice-contingent[11,43]. The hot stove effect can cause a preference

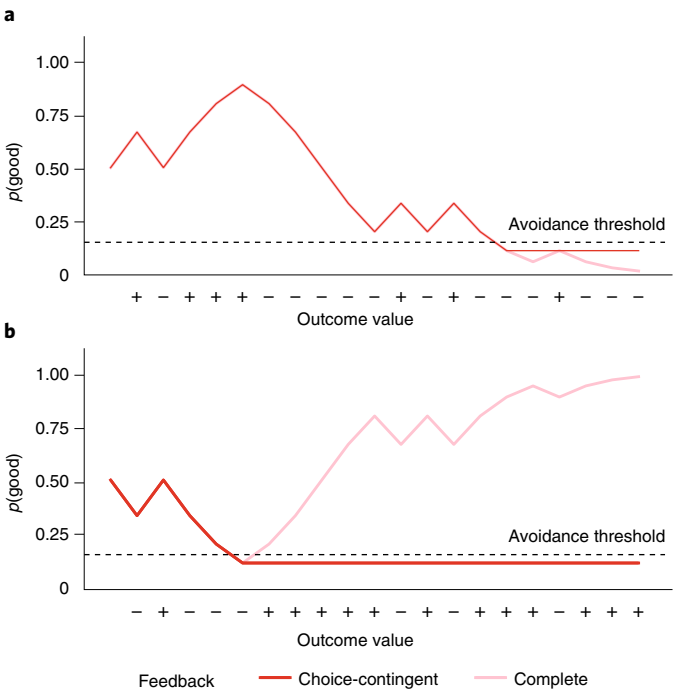

Fig. 2 | An agent's beliefs about whether an option is mostly good or mostly bad evolve as the agent experiences a series of positive and negative outcomes, potentially causing the hot stove effect. The data are based on a simple Bayesian learning algorithm and outcomes are encountered left to right. If the agent's belief dips below a certain threshold, the agent avoids the item or option for all future encounters and receives no further feedback about its value. **a**, The option is usually bad but seems deceptively good at first. The agent continues choosing the item until eventually discovering its belief was wrong. **b**, The option is usually good but seems bad at first. On the basis of early experience, the agent avoids the item, exhibiting the hot stove effect and missing out on later rewards. The agent's beliefs about the item diverge from what they would be if the agent received complete feedback regardless of its choices.

for novel alternatives over known ones, because the valuations of novel options have not yet been influenced by the effect[44]. In situations where multiple features of an option are revealed only after choosing it, such as how a person's intelligence and sense of humour are revealed through conversation, the hot stove effect can lead to a form of illusory correlation[45]. Importantly, the hot stove effect is not limited to human behaviour. It is also exhibited by non-human decision makers, such as foraging honey bees[46], and has been shown to emerge even in an optimal decision-making agent[47].

Although the classic hot stove effect applies to simple, discrete actions or options, research has shown how similar effects can emerge when a decision maker learns about and generalizes across options with multiple features, as is often the case in human and machine learning. For instance, when hiring for a company, a recruiter will not repeatedly hire or reject a single applicant to learn about them from experience. Instead, the recruiter will make decisions about many applicants, each of whom has many attributes, and will gradually learn the features that produce the best fit for the company over the course of many hires. In this type of situation, a negative outcome can cause avoidance not only of the experienced option, but also of similar options, even when the similar options are actually positive. Over time, this interaction between the classic hot stove effect and people's ability to generalize can cause the decision maker to believe a greater space of options is negative than is truly the case. This phenomenon is seen both in human behaviour in the laboratory[48] and in simulations with simple neural network models[49].

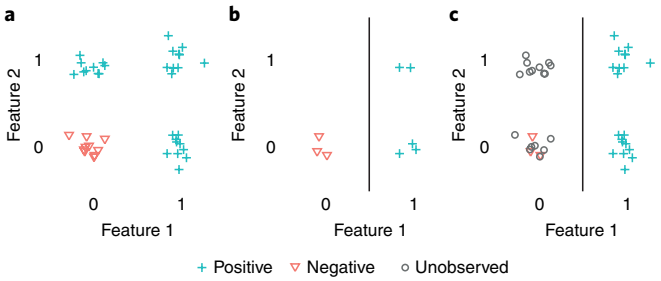

Fig. 3 | An 'attentional learning trap' can emerge with choice-contingent feedback in some environments. a, In the true structure of this environment, the presence of either of two features produces a positive outcome. **b**, On the basis of an agent's early experience, only a single feature may appear important, causing the agent to form a one-dimensional decision boundary, shown by the black vertical line. This is especially likely if the agent prefers simple representations, as is true of humans and many learning algorithms. **c**, Once the decision boundary is learned, items on the negative side of it are avoided, and error-correcting feedback about these items that would produce a more complex but correct boundary is never received.

Recent work has shown how choice-contingent feedback can also cause a learner to ignore important features of the options[50]. This bias, termed the attentional learning trap, can occur because people, like many machine learning algorithms, prefer to learn a simple, low-dimensional representation of the environment[37,51]. Returning to the recruiting example, suppose that the recruiter reviews applicants with and without bachelor's degrees and with and without two years of relevant experience, and that applicants with either qualification would perform well at the company (shown schematically in Fig. 3a). After hiring a few individuals, the recruiter might observe that all applicants with a bachelor's degree are performing well, while some without are not, and thus hypothesize that this is the sole important feature (Fig. 3b). If the recruiter acts on the basis of this hypothesis, then in the future they will hire applicants with bachelor's degrees and reject those without, and thus never learn that applicants with two years of experience also perform well, even when they lack a bachelor's degree (Fig. 3c). The choice-contingency of applicant feedback traps the recruiter into paying attention to only one dimension of the applications, which leaves them rejecting suitable applicants and having false beliefs about the attributes an applicant should have.

The choice-contingent learning problems faced by AI systems are similar to, and often match exactly, those faced by humans. The recruiter screening applications at a large company, for instance, could easily be a machine, and not a human. In some settings where machines are beginning to have a role, the stakes are even higher than whether an applicant lands a job. Among the highest-stakes settings are those in criminal justice. Several research groups have investigated the use of algorithms to help judges decide which defendants should have bail set before trial and which should be released without bail[38,52]. The goal is to set bail only for defendants who would not otherwise appear at their court date; however, because those defendants who are unable to afford the bail set for them wait in jail until their trial, there is no way to ascertain whether they would have appeared had they been free, and thus the classic challenges of choice-contingent feedback emerge.

A related area in which the effect of choice-contingent feedback has received particular attention is predictive policing. Algorithmic approaches that send officers to predicted high-crime locations have shown apparent success in reducing crime[8]. These algorithms work by predicting that crime will occur near where it has been observed in the past. However, critics of predictive policing have pointed out that, because crime is more likely to be observed when there are many police there to observe it, patterns of observed crime can diverge from the pattern of actual crime[9]. Much as a person learns to avoid a coworker after an initial bad impression, the algorithm learns not to send officers to areas it believes to be low-crime and to focus on 'high-crime' areas. Over time, a feedback loop can develop, in which more and more police can be sent to a small number of crime hot spots, thereby causing a greater and greater percentage of crime to be observed there[10].

Owing in part to the high potential impact of bias in these types of domain, machine learning researchers have begun to study how to mitigate the effects of choice-contingent feedback. One recent paper developed the concept of 'fair reinforcement learning', which requires that an algorithm never prefers an inferior option to a higher-payoff option, and introduced algorithms for approximately fair reinforcement learning[53]. Work on bail decisions has used sophisticated statistical techniques to infer the probable outcome of release for defendants who were held before trial[38,52]. In practice, simpler approaches may be used, such as allowing an algorithm to learn only from a sample of data for which it had no role in the choices made and feedback received (for example, a set of job applicants in which all candidates were accepted or candidates were accepted randomly)[54].

Although all of these approaches are useful, their limits are demonstrated by theoretical work on the hot stove effect and related biases. One critique of predictive policing algorithms focuses on the fact that they do not account for the effect of policing rates on observed crime rate, and shows how adjusting for policing rate can prevent feedback loops[10]. What research on the hot stove effect reveals is the trade-off required to stop this bias from occurring. As demonstrated by Le Mens and Denrell[47], the hot stove effect is a property of even an optimal decision-making agent that seeks to maximize reward. To avoid the effect, an agent must engage in sub-optimal behaviour, decreasing its long-term reward to collect additional information. In other words, although the hot stove effect can be reduced to some degree by making algorithms 'better', eliminating it completely requires acknowledging other goals beyond optimizing performance in a narrow sense. In the case of policing, this means accepting that fewer crimes may be detected in the interest of ensuring that crime has a more equitable chance of being detected at different locations. Even though this trade-off may be worthwhile, deciding whether to make it and to what extent are challenging policy decisions, as illustrated by the tensions faced in fairly applying algorithms to criminal justice decisions[39].

Aside from these theoretical and policy concerns, the psychological literature on choice-contingent feedback biases may have practical lessons for designers of machine learning systems, particularly in specifying where these biases could occur. Practitioners should consider the potential for the hot stove effect whenever some options produce higher-variance rewards than others[42], some options are newer than others[44], or the model received choice-contingent feedback for some options but not others[11,43]. Where the hot stove effect has been found, it can be reduced by using reinforcement learning algorithms[16] to explore under-sampled options; alternatively, simpler methods such as training on a non-choice-contingent dataset[54] may be used to reduce it. In cases where options have multiple features, the ways in which bias may emerge are less clear-cut. Although psychological studies have shown ways in which simple algorithms can generalize incorrectly in this domain[49,50], little is known about how these sorts of biases scale up to applied machine learning domains, where both the dimensionality of the feature space and the generalization abilities of the algorithms are much greater. This is an area that carries both a high risk of unseen biases and a high potential reward from further research.

Biases from within

Many of the biases discussed in the previous sections are ecological in nature, caused not mainly by a particular mechanism in the

human mind but by the kind of data produced by the environment and the individual's interaction with it[55]. This characteristic makes the links to machine learning particularly clear, because the decision setting can remain the same even as the decider changes. However, much of psychology is more internally oriented, and many of the best known examples of human bias, such as those identified by Kahneman and Tversky as part of their 'biases and heuristics' programme of research, seem to be side effects of our internal mental architecture, decision processes and inconsistent preferences[20,56]. At first glance, these types of bias seem less relevant to artificial systems. But the extensive research into the causes of these biases, particularly efforts to understand how they may be rational in certain settings or have computational justifications, has shown that here too there are valuable lessons for the development of AI.

Take, for instance, one of the crowning achievements of the field of human judgement and decision-making: prospect theory[56], which describes how humans make choices among risky alternatives. Whereas economic theory states that a rational agent should have stable risk preferences—either risk-seeking or risk averse—people's risk preferences change relative to their reference point, forming an S-shaped utility curve, as shown in Fig. 4a. When outcomes appear as gains relative to their current status, people are risk-averse, whereas when outcomes appear as losses, people are risk-seeking. For example, a person might prefer the certainty of winning US$1 over flipping a coin for a 50% chance of winning US$2, but simultaneously prefer a 50% chance of losing US$2 over the certainty of losing US$1.

Prospect theory seems like a quirk unique to the human mind, and indeed there are many situations where people exhibit clearly irrational risk preferences[56] that are easily avoidable by a machine. But work has also shown that in some situations reference-dependent risk preferences can be a side effect of Bayesian inference with strong priors. Specifically, suppose that the risks and rewards of two options are not observed through explicit description (as in a prospectus for a mutual fund) but as a set of past experiences. In this case, the values of the options may be estimated in a Bayesian manner, combining the observed past outcomes with a prior belief distribution that is likely to be centred at the current reference point. With certain specifications of the internal model, the expected values of more variable (and thus more risky) options are drawn towards the reference point more than are those of less variable options, creating a preference for consistent options in the domain of gains and risky options in the domain of losses[57] (Fig. 4b).

Since these reference-dependent risk preferences are the consequences of rational inference, they are not necessarily harmful biases that should be corrected. However, they can affect various algorithmic decisions, especially when decisions vary in their relation to a reference point. For example, in a part of town with bad restaurants that are all below a town-wide reference point, a recommendation algorithm may suggest an option with many one- and three-star reviews over one with mostly two-star reviews; by contrast, in a part of town with good restaurants, the same algorithm may suggest restaurants with many four-star reviews over those with a mix of three- and five-star reviews. Whether this sort of pattern could constitute, in more consequential applications, a harmful bias that would need to be corrected requires further examination.

Other findings from Kahneman's and Tversky's biases and heuristics research have also been analysed in terms of Bayesian reasoning. However, these phenomena have been explained not as rational but as side effects of approximate inference under computational resource constraints. For instance, the anchoring bias, in which people's estimates of an unknown quantity are shifted towards a value that is initially provided or comes to mind easily, can be caused by Markov chain sampling with insufficient time for convergence[58]. The availability heuristic, in which the probabilities of rare but easily recalled events are over-weighted, could emerge from a form of utility-weighted importance sampling[59]. This is an interesting case in which algorithmic ideas from machine learning have provided inspiration for psychological theorizing, and it is worth considering whether the resulting psychological insights may be brought back to machine learning. The resource constraints needed to produce these biases are extreme, and may be rare in practice. However, other forms of approximate inference that are used in practice, such as variational inference, do affect model performance in ways that are not yet well understood[60,61]. These case studies from psychology may be useful guides to machine learning researchers as they seek to achieve effective inference under computational resource constraints[62].

While the biases and heuristics literature has often focused on how human decision-making mechanisms can lead to poor outcomes, a separate and at times competing strand of research has focused on how internal heuristics can lead to very accurate decisions, terming them 'fast and frugal' heuristics[63,64]. Heuristics that perform surprisingly well include tallying feature values and 'take-the-best', in which the feature values of two options are compared from most to least important feature until one option wins on a feature comparison[65]. Researchers have outlined the types of

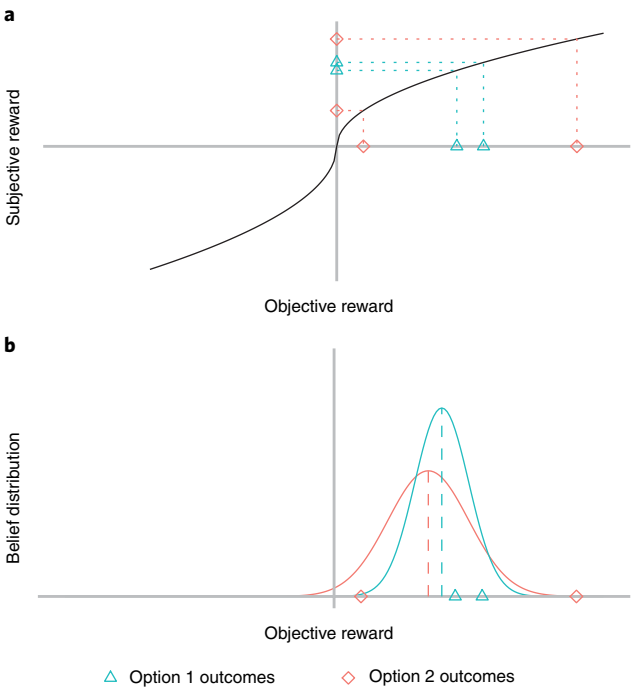

Fig. 4 | Reference-dependent risk preferences can be produced by Bayesian prediction. In both panels, two options are displayed, each with two outcomes that have been observed once. The options have the same mean, but option 2 is higher risk because the two observed outcomes are more different from each other. **a**, The classic explanation of reference-dependent risk preferences, with an S-shaped utility function mapping objective reward to subjective reward. The bend in the function causes the low-risk option to have higher mean reward in the subjective space, causing risk aversion. **b**, An alternative, Bayesian pathway to this effect, assuming the distributions of rewards have unknown mean and variance and the prior beliefs of the reward distribution means are centred near zero. The posterior belief distributions, plotted for each option with a dashed line at the posterior mean, show that the the low-risk option has a lower inferred variance and a higher inferred mean, causing risk aversion. For both causes of risk aversion, the direction of the effect flips when the options are switched from positive outcomes (gains) to negative ones (losses), causing risk seeking in this domain.

environment in which tallying and take-the-best do and do not perform well[66,67], and have demonstrated that these heuristics too might have a rational formalization—in this case, as Bayesian regression with extreme priors[68]. This extensive literature may become more important as the machine learning field seeks to create simpler and more interpretable models[69]. An algorithm for bail decisions has already been developed using a transparent tallying-like procedure[52] that is nearly as effective as a random forest model and similar to models developed decades ago by psychologists[41]. As the creation of alternatives to complex, opaque models continues, existing research into fast and frugal heuristics can specify when these alternatives will work and how they relate to more traditional machine learning techniques.

Conclusion

Science and technology often advance through inspiring metaphors. Some of the recent interest in machine learning and AI stems precisely from the comparison between machines and humans, and the idea that machine-based systems implement aspects of human cognition but improve on human abilities. However, a more nuanced version of this idea includes the acknowledgement that these systems, as 'intelligent' as they are, fall victim to the same traps and hiccups that people do. A healthy attitude towards recent advances in AI would be to recognize that rather than being free of bias, certain biases are likely to be fundamental to what it means to be an intelligent adaptive agent operating in a vague and uncertain world.

This Perspective has sought to highlight a few areas in which machine learning and AI practitioners can learn from the psychology of learning and decision-making biases, but the examples discussed here are far from exhaustive. Melioration, in which an incomplete representation of the environment can trap a greedy decision maker into persistent poor choices, is a vital concept for reinforcement learning models[70,71]. Categorical perception, which allows people to differentiate subtle perceptual categories, can come at the cost of being able to distinguish unfamiliar categories such as non-native-speech sounds[72,73]. It may therefore be of interest to deep learning researchers in speech recognition and vision[2]. Finally, numerous sampling biases, including those related to survivorship-based sampling and censored sampling, have been uncovered or elucidated by researchers at the intersection of psychology and management science[74–77].

It is striking that questions about biased learning and decision-making that have interested psychologists since the 1960s, such as why the properties of rare groups are often misestimated[29,78], are a focus of machine learning research today[36,40]. In most cases, the conclusions reached by psychologists about human biases do not directly translate into an improved machine learning model. The technical solutions to algorithmic bias will probably come from more computational fields, just as neural network and reinforcement learning research have flourished in computer science departments despite their roots in psychology. But what the psychology literature can provide is a roadmap to where and how biases may pop up, the tradeoffs that can be made to reduce them, and the situations where biased but simple decision-making may be harmless or beneficial. As machine learning and AI systems move into areas where they are making life-changing decisions, such as healthcare and criminal justice, this roadmap is sorely needed[4]. It is our hope that the study of what Amos Tversky called 'natural stupidity' can lead to more effective and equitable AI.

Received: 6 February 2019; Accepted: 8 March 2019;
Published online: 9 April 2019

References

1. Silver, D. et al. Mastering the game of Go without human knowledge. *Nature* **550**, 354–359 (2017).
2. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
3. Lazer, D., Kennedy, R., King, G. & Vespignani, A. The parable of Google flu: traps in big data analysis. *Science* **343**, 1203–1206 (2014).
4. Campolo, A., Sanfilippo, M., Whittaker, M. & Crawford, K. A. I. *Now 2017 Report* (AI Now Institute, 2017); https://ainowinstitute.org/AI_Now_2017_Report.pdf.
5. O'Neil, C. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (Broadway Books, 2017).
6. Barocas, S. & Selbst, A. Big data's disparate impact. *Calif. Law Rev.* **104**, 671–729 (2016).
7. Buolamwini, J. & Gebru, T. Gender shades: intersectional accuracy disparities in commercial gender classification. In *Proc. Mach. Learn. Res.* Vol. 81 (eds Friedler, S. A, & Wilson, C.) 77–91 (PMLR, 2018).
8. Mohler, G. O. et al. Randomized controlled field trials of predictive policing. *J. Am. Stat. Assoc.* **110**, 1399–1411 (2015).
9. Lum, K. & Isaac, W. To predict and serve? *Significance* **13**, 14–19 (2016).
10. Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C. & Venkatasubramanian, S. Runaway feedback loops in predictive policing. In *Proc. Mach. Learn. Res.* Vol. 81 (eds Friedler, S. A. & Wilson, C.) 1–12 (PMLR, 2018).
11. Denrell, J. Why most people disapprove of me: experience sampling in impression formation. *Psychol. Rev.* **112**, 951–978 (2005).
12. Crawford, K. The hidden biases in big data. *Harvard Business Review* https://hbr.org/2013/04/the-hidden-biases-in-big-data (1 April 2013).
13. Hertwig, R., Barron, G., Weber, E. U. & Erev, I. Decisions from experience and the effect of rare events in risky choice. *Psychol. Sci.* **15**, 534–539 (2004).
14. Wiener, N. *Cybernetics: Or Control and Communication in the Animal and the Machine* (MIT Press, 1948).
15. von Neumann, J. *The Computer and the Brain* (Yale Univ. Press, 1958).
16. Sutton, R. S. & Barto, A. G. *Reinforcement Learning: An Introduction* (Cambridge Univ. Press, 1998).
17. Rumelhart, D. E., McClelland, J. L. & PDP Research Group. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations* (MIT Press, 1986).
18. Marcus, G. Deep learning: a critical appraisal. Preprint at https://arxiv.org/abs/1801.00631 (2018).
19. Lake, B. M., Ullman, T. D., Tenenbaum, J. B. & Gershman, S. J. Building machines that learn and think like people. *Behav. Brain Sci.* **40**, 1–101 (2016).
20. Tversky, A. & Kahneman, D. Judgment under uncertainty: heuristics and biases. *Science* **185**, 1124–1131 (1974).
21. Lake, B. M., Salakhutdinov, R. & Tenenbaum, J. B. Human-level concept learning through probabilistic program induction. *Science* **350**, 1332–1338 (2015).
22. Chomsky, N. *Aspects of the Theory of Syntax* (MIT Press, 1965).
23. Fox, C. R. & Hadar, L. Decisions from experience' = sampling error + prospect theory: reconsidering Hertwig, Barron, Weber & Erev (2004). *Judgm. Decis. Mak.* **1**, 159–161 (2006).
24. Harris, A. J. & Hahn, U. Unrealistic optimism about future life events: a cautionary note. *Psychol. Rev.* **118**, 135–154 (2011).
25. Chambers, J. R., Windschitl, P. D. & Suls, J. Egocentrism, event frequency, and comparative optimism: when what happens frequently is 'more likely to happen to me'. *Personal. Social. Psychol. Bull.* **29**, 1343–1356 (2003).
26. Teodorescu, K. & Erev, I. On the decision to explore new alternatives: the coexistence of under- and over-exploration. *J. Behav. Decis. Mak.* **27**, 109–123 (2014).
27. Fuller, R. Behavior analysis and unsafe driving: warning—learning trap ahead! *J. Appl. Behav. Anal.* **24**, 73–75 (1991).
28. Szollisi, A., Liang, G., Konstantinidis, E., Donkin, C. & Newell, B. R. Simultaneous underweighting and overestimation of rare events: unpacking a paradox. *J. Exp. Psychol. Gen.* (in the press).
29. Hamilton, D. L. & Gifford, R. K. Illusory correlation in interpersonal perception: a cognitive basis of stereotypic judgments. *J. Exp. Social. Psychol.* **12**, 392–407 (1976).
30. Mullen, B. & Johnson, C. Distinctiveness-based illusory correlations and stereotyping: a meta-analytic integration. *Br. J. Social. Psychol.* **29**, 11–28 (1990).
31. Eder, A. B., Fiedler, K. & Hamm-Eder, S. Illusory correlations revisited: the role of pseudocontingencies and working-memory capacity. *Q. J. Exp. Psychol.* **64**, 517–532 (2011).
32. Kutzner, F., Vogel, T., Freytag, P. & Fiedler, K. A robust classic: illusory correlations are maintained under extended operant learning. *Exp. Psychol.* **58**, 443–453 (2011).
33. Fiedler, K. & Kutzner, F. in *The Wiley Blackwell Handbook of Judgment and Decision Making* (eds Keren, G. & Wu, G.) 380–403 (Wiley, 2015).
34. Fiedler, K., Walther, E., Freytag, P. & Plessner, H. Judgment biases in a simulated classroom — a cognitive – environmental approach. *Organ. Behav. Human. Decis. Process.* **88**, 527–561 (2002).
35. Halevy, A., Norvig, P. & Pereira, F. The unreasonable effectiveness of data. *IEEE Intell. Syst.* **24**, 8–12 (2009).
36. Lerman, J. Big data and its exclusions. *Stanf. Law Rev.* **66**, 55–63 (2013).
37. Tibshirani, R. Regression selection and shrinkage via the lasso. *J. R. Stat. Soc. Ser. B* **58**, 267–288 (1996).

38. Kleinberg, J., Mullainathan, S. & Raghavan, M. Inherent trade-offs in the fair determination of risk scores. Preprint at https://arxiv.org/abs/1609.05807 (2016).
39. Corbett-Davies, S., Pierson, E., Feller, A., Goel, S. & Huq, A. Algorithmic decision making and the cost of fairness. In *Proc. 23rd Conf. Knowledge Discovery and Data Mining*, 797-806 (2017).
40. Chen, I., Johansson, F. D. & Sontag, D. Why is my classifier discriminatory? *Adv. Neural Inf. Process. Syst.* **31**, 3543–3554 (2018).
41. Dawes, R. The robust beauty of improper linear models in decision making. *Am. Psychol.* **34**, 571–582 (1979).
42. Denrell, J. & March, J. G. Adaptation as information restriction: the hot stove effect. *Organ. Sci.* **12**, 523–538 (2001).
43. Liu, C., Eubanks, D. L. & Chater, N. The weakness of strong ties: sampling bias, social ties, and nepotism in family business succession. *Leadersh. Q.* **26**, 419–435 (2015).
44. Le Mens, G., Kareev, Y. & Avrahami, J. The evaluative advantage of novel alternatives: an information-sampling account. *Psychol. Sci.* **27**, 161–168 (2016).
45. Denrell, J. & Le Mens, G. Seeking positive experiences can produce illusory correlations. *Cognition* **119**, 313–324 (2011).
46. Niv, Y., Joel, D., Meilijson, I. & Ruppin, E. Evolution of reinforcement learning in uncertain environments: a simple explanation for complex foraging behaviors. *Adapt. Behav.* **10**, 5–24 (2002).
47. Le Mens, G. & Denrell, J. Rational learning and information sampling: on the 'naivety' assumption in sampling explanations of judgment biases. *Psychol. Rev.* **118**, 379–392 (2011).
48. Fazio, R. H., Eiser, J. R. & Shook, N. J. Attitude formation through exploration: valence asymmetries. *J. Personal. Social. Psychol.* **87**, 293–311 (2004).
49. Eiser, J. R., Fazio, R. H., Stafford, T. & Prescott, T. J. Connectionist simulation of attitude learning: asymmetries in the acquisition of positive and negative evaluations. *Personal. Social. Psychol. Bull.* **29**, 1221–1235 (2003).
50. Rich, A. S. & Gureckis, T. M. The limits of learning: exploration, generalization, and the development of learning traps. *J. Exp. Psychol. Gen.* **147**, 1553–1570 (2018).
51. Shepard, R. N., Hovland, C. L. & Jenkins, H. M. Learning and memorization of classifications. *Psychol. Monogr.* **75**, 1689–1699 (1961).
52. Jung, J., Concannon, C., Shroff, R., Goel, S. & Goldstein, D. G. Simple rules for complex decisions. *Harvard Business Review* https://hbr.org/2017/04/creating-simple-rules-for-complex-decisions (19 April 2017).
53. Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J. & Roth, A. Fairness in reinforcement learning. In *Proc. Mach. Learn. Res.* Vol. 71 (eds Precup, D. & Whye Teh, Y.) 1617–1626 (PMLR, 2017).
54. Sculley, D. et al. Hidden technical debt in machine learning systems. *Adv. Neural Inf. Process. Syst.* **28**, 2503–2511 (2015).
55. Fiedler, K. Beware of samples! A cognitive-ecological sampling approach to judgment biases. *Psychol. Rev.* **107**, 659–676 (2000).
56. Kahneman, D. & Tversky, A. Prospect theory: an analysis of decision under risk. *Économ. J. Econom. Soc.* **47**, 263–292 (1979).
57. Denrell, J. Reference-dependent risk sensitivity as rational inference. *Psychol. Rev.* **122**, 461–484 (2015).
58. Lieder, F., Griffiths, T. L. & Goodman, N. D. Burn-in, bias, and the rationality of anchoring. *Adv. Neural Inf. Process. Syst.* **25**, 2790–2798 (2012).
59. Lieder, F., Griffiths, T. L. & Hsu, M. Overrepresentation of extreme events in decision making reflects rational use of cognitive resources. *Psychol. Rev.* **125**, 1–32 (2018).
60. Blei, D. M., Kucukelbir, A. & McAuliffe, J. D. Variational inference: a review for statisticians. *J. Am. Stat. Assoc.* **112**, 859–877 (2017).
61. Blei, D. M. & Lafferty, J. D. Dynamic topic models. In *Proc. 23rd Int. Conf. Machine Learning* 113–120 (ACM, 2006).
62. Gershman, S. J., Horvitz, E. J. & Tenenbaum, J. B. Computational rationality: a converging paradigm for intelligence in brains, minds, and machines. *Science* **349**, 273–278 (2015).
63. Gigerenzer, G. & Brighton, H. Homo heuristicus: why biased minds make better inferences. *Top. Cogn. Sci.* **1**, 107–143 (2009).
64. Katsikopoulos, K. V. Bounded rationality: the two cultures. *J. Econ. Methodol.* **21**, 361–374 (2014).
65. Czerlinski, J., Gigerenzer, G. & Goldstein, D. G. How Good Are Simple Heuristics? in *Simple Heuristics That Make Us Smart* (eds Gigerenzer, G., Todd, P.M., & The ABC Research Group) 97-118 (Oxford Univ. Press, 1999).
66. Martignon, L. & Hoffrage, U. in *Simple Heuristics That Make Us Smart* (eds Gigerenzer, G., Todd, P. M., & The ABC Research Group) 119–140 (Oxford Univ. Press, 1999).
67. Hogarth, R. M. & Karelaia, N. 'Take-the-best' and other simple strategies: why and when they work 'well' with binary cues. *Theory Decis.* **61**, 205–249 (2006).
68. Parpart, P., Jones, M. & Love, B. C. Heuristics as Bayesian inference under extreme priors. *Cogn. Psychol.* **102**, 127–144 (2018).
69. Doshi-Velez, F. & Kim, B. Towards a rigorous science of interpretable machine learning. Preprint at https://arxiv.org/abs/1702.08608 (2017).
70. Herrnstein, R. J. & Prelec, D. Melioration: a theory of distributed choice. *J. Econ. Perspect.* **5**, 137–156 (1991).
71. Gureckis, T. M. & Love, B. C. Short-term gains, long-term pains: how cues about state aid learning in dynamic environments. *Cognition* **113**, 293–313 (2009).
72. Polka, L. & Werker, J. F. Developmental changes in perception of nonnative vowel contrasts. *J. Exp. Psychol. Human. Percept. Perform.* **20**, 421–435 (1994).
73. Goldstone, R. Influences of categorization on perceptual discrimination. *J. Exp. Psychol. Gen.* **123**, 178–200 (1994).
74. Levinthal, D. A. & March, J. G. The myopia of learning. *Strateg. Manag. J.* **14**, 95–112 (1993).
75. Denrell, J. Vicarious learning, undersampling of failure, and the myths of management. *Organ. Sci.* **14**, 227–243 (2003).
76. Feiler, D. C., Tong, J. D. & Larrick, R. P. Biased judgment in censored environments. *Manag. Sci.* **59**, 573–591 (2013).
77. Hogarth, R. M., Lejarraga, T. & Soyer, E. The two settings of kind and wicked learning environments. *Curr. Dir. Psychol. Sci.* **24**, 379–385 (2015).
78. Chapman, L. J. Illusory correlation in observational report. *J. Mem. Lang.* **6**, 151–155 (1967).

Competing interests

A.S.R. is employed by Flatiron Health, an independent subsidiary of Roche.

Additional information

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence should be addressed to A.S.R.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.