Alterations in choice behavior by manipulations of world model

C. S. Green[a,b,1], C. Benson[a], D. Kersten[a,b,c], and P. Schrater[a,b,d]

[a]Department of Psychology, [b]Center for Cognitive Science, and [d]Department of Computer Science, University of Minnesota, Minneapolis, MN 55455; and [c]Max Planck Institute for Biological Cybernetics, 72012 Tübingen, Germany

Edited* by Wilson S. Geisler, University of Texas, Austin, TX, and approved July 22, 2010 (received for review February 10, 2010)

How to compute initially unknown reward values makes up one of the key problems in reinforcement learning theory, with two basic approaches being used. Model-free algorithms rely on the accumulation of substantial amounts of experience to compute the value of actions, whereas in model-based learning, the agent seeks to learn the generative process for outcomes from which the value of actions can be predicted. Here we show that (*i*) "probability matching"—a consistent example of suboptimal choice behavior seen in humans—occurs in an optimal Bayesian model-based learner using a max decision rule that is initialized with ecologically plausible, but incorrect beliefs about the generative process for outcomes and (*ii*) human behavior can be strongly and predictably altered by the presence of cues suggestive of various generative processes, despite statistically identical outcome generation. These results suggest human decision making is rational and model based and not consistent with model-free learning.

decision making | probability matching | reinforcement learning

Given a limited set of data about the world, what is the best thing to do? This question lies at the heart of all decision making, from simple everyday errands to elaborate and complex scientific experiments. If the reward amount for each possible action is known in advance, it is a straightforward process to make choices that maximize reward. In the real world, however, reward values are nearly always initially unknown and computing them is not trivial. Thus, understanding how to learn and compute reward is one of the key problems in reinforcement learning theory. Computing the optimal policy (i.e., determining the "best thing to do") requires acquiring one of two types of knowledge. In *model-free* learning, an agent must accumulate a substantial amount of experience regarding the consequences of taking various actions in various states, from which the average value of the states can be learned. In *model-based* learning, an agent must acquire a "world model," which constitutes beliefs about how the world generates outcomes in response to actions. Although both model-free and model-based reinforcement-learning algorithms have been the subject of much study in computer science and machine learning, model-free algorithms have been primarily used as models of human choice behavior.

Whereas it is clear that our survival depends on the ability to make appropriate decisions from incomplete and ambiguous information, numerous studies in economics, psychology, and neuroscience have consistently found highly suboptimal behavior in seemingly simple decision tasks. Why is this? Consider the sequential binary decision task, which involves a choice between two options, one with a higher probability of success than the other (e.g., 70% vs. 30% of trials). The optimal strategy for this task is to determine which option has a higher probability of success and then choose only that option. Humans, however, tend to sample the alternatives in proportion to the options' respective probabilities of being correct. This is an exceptionally consistent effect known as "probability matching." It has been replicated in dozens of laboratories, under myriad task conditions, and is extremely robust, persisting for thousands of trials (1–15). Most theories treat this behavior as a fundamental failure of rational decision making. In contrast, we propose that humans typically engage in rational decision making arising from model-based, rather than model-free learning. From this perspective, nonoptimal decision making, such as probability matching, emerges as a consequence of a poor match between the model used by human subjects to interpret the data and the generative model used in a typical experiment.

Here we show that (*i*) probability matching behavior occurs in even an optimal Bayesian model-based learner that is initialized with ecologically plausible, but incorrect beliefs about the generative process for outcomes and (*ii*) human behavior can be strongly and predictably altered by the presence of cues suggestive of various generative processes, despite statistically identical outcome generation. These results suggest human decision making is rational and model based and not consistent with model-free learning.

In the sequential binary decision task, the goal is to make choices that maximize the number of successes. There are three key facts about the generative process for outcomes that the subject must learn or infer to compute the optimal strategy. First, the probability of success is greater for one of the two options (e.g., choosing "option A" leads to success 70% of the time, whereas choosing "option B" leads to success 30% of the time). Second, outcomes are independent across time (i.e., if option A was successful on the previous trial, it does not increase or decrease the probability that option A will be successful on the next trial). And third, the outcomes are coupled, or, in other words, something can be inferred about one option's reward probability from observing the other. Given this model for outcome generation, the normative optimal strategy is simply to always choose the option with the higher probability of success.

Although the true generative model in the binary choice task is one in which outcomes are temporally independent and coupled, this is not necessarily the most ecologically plausible generative process. Thus, it would be surprising if subjects naturally posit such a model. Instead, we suggest that subjects may initially consider a model in which outcomes are temporally *dependent* and *uncoupled*. As a purely illustrative example, imagine a person who is faced with the choice between hunting and gathering. On day 1 he chooses to gather and is pleased to find an orchard bearing ripe fruit ("a success"). On day 2, he reasons that because there was ripe fruit in the orchard yesterday, it is likely that there is ripe fruit there again today and thus the probability of success if he chooses to gather is high (temporal dependence). It is also intuitive that his success at gathering yesterday taught him nothing about what *would have happened*, had he attempted to hunt instead of gather (uncoupled outcomes). The two key insights should be apparent: (*i*) Processes that generate outcomes in the external world tend to change slowly and thus the outcomes that they generate have a high degree of temporal dependence and (*ii*) it is rarely the case

Author contributions: C.S.G., D.K., and P.S. designed research; C.S.G., C.B., and P.S. performed research; C.S.G. and P.S. analyzed data; and C.S.G., D.K., and P.S. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

[1]To whom correspondence should be addressed. E-mail: csgreen@umn.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1001709107/-/DCSupplemental.

that the outcome of one's actual choice offers any real information about what would have happened had another choice been made.

Results

Probability Matching in an Optimal Bayesian Learner. According to our hypothesis, trial-by-trial errors in decision making such as probability matching are a simple consequence of a rational agent initializing learning with a more general, but incorrect model of the world, which biases the way it interprets outcomes and selects actions (Fig. 1*A*). In Fig. 1 *B* and *C*, we show that human-like errors in decision making can emerge from an optimal Bayesian learner initialized with the erroneous, but ecologically plausible assumptions mentioned above (temporal dependence, uncoupled outcomes—*Methods* and *SI Appendix S1* and *S2*).

In addition to exhibiting probability matching (Fig. 1*B*), the Bayesian model shows other trends characteristic of human behavior. Of particular note is that the cumulative histogram of choice run lengths is well fit by a power function (Fig. 1*C*) that demonstrates a clear temporal dependence in choices (i.e., subjects make more and longer runs of the same choice than would be predicted if choices were temporally independent; see also refs. 16 and 17 for examples of temporal structure in choice behavior). Critically, these behaviors emerge despite the fact that the model employs a strictly maximizing decision rule (i.e., the agent chooses the option with the highest expected value on every trial). This is in contrast with other current models in the field that produce human-like behavior via the use of stochastic (soft-max) decision rules along with parameters that enforce a certain degree of "stickiness" in choices (18). The model results are also consistent with the known fact that probability-matching behavior in humans can last for thousands of trials (13), as the pattern of observed outcomes does not allow the initial assumptions to be quickly "unlearned." It is important to note that this model is not meant to provide a trial-by-trial account of human behavior. Instead, it is simply meant to demonstrate that a fully "rational" agent can produce seemingly "irrational" behavior as a simple consequence of using an incorrect world model. And although these results certainly do not *prove* that humans are making the assumptions initially provided to the model (and in fact we believe it is likely that subjects posit a range of possible temporal dependence models running the gamut between state persistence and state transience), they suggest that human choice behavior should be determined not just by the outcomes that are observed, but also by the model of the world used to interpret the outcomes.

Experimental Manipulations of World Model. Given our hypothesis it should be possible to provide experimental cues that alter the most likely world model and thus substantially affect subject behavior without changing the outcome statistics themselves. Of particular interest to us was to compare choice behavior in tasks analogous to those used throughout the binary choice literature, but where we alter the environment in such a way as to be suggestive of different generative models for the task. To this end, we embedded a standard sequential binary choice task within two environments that

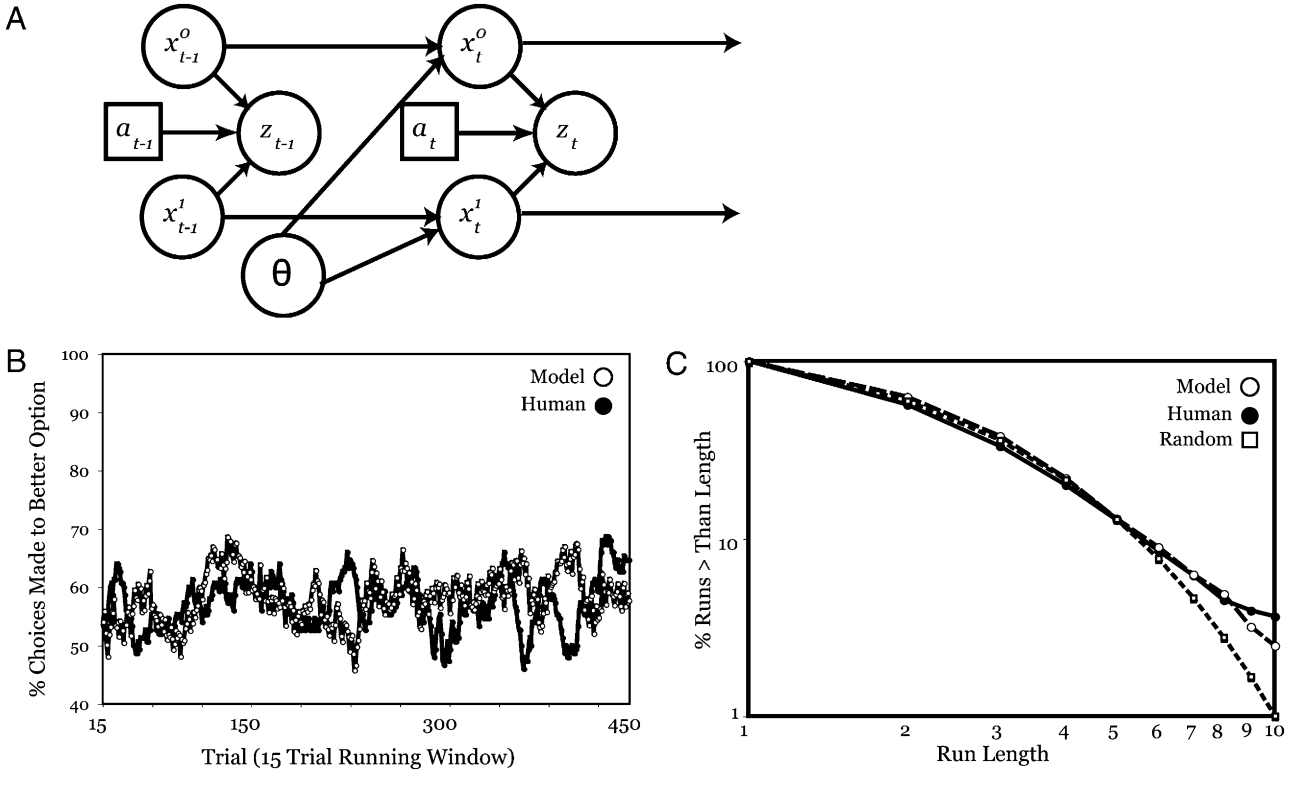

Fig. 1. (*A*) Graphical representation of task. At each time step the two choice options (x^0 and x^1) are in one of two states (win/lose). However, the subject is only allowed to observe the state (z) of one of the options through their choice (a). After each choice, the options have some probability of transitioning from their current state (win/lose) to the opposite state, according to a transition matrix θ (which implicitly contains information regarding the overall success probabilities of the two options, whether or not the options are coupled, as well as the degree of temporal independence of outcomes over time) (*Methods*). (*B* and *C*) Choice behavior in humans and optimal Bayesian agent with incorrect world model. (*B*) The percentage of choices to the better target in 15-trial moving windows in humans and Bayesian agent. Qualitatively similar probability matching behavior (the better option is successful in 60% of trials in this task) is seen in both humans and an optimal Bayesian agent that is initialized with an incorrect world model. (*C*) The percentage of runs that are of the given length or longer (e.g., – ≈60% of runs are ≥2 trials and ≈4% of runs are ≥10 trials). Both humans and the Bayesian agent show a function that is linear in log–log coordinates (power law), which indicates temporal dependence in choices (more long runs than would be expected if subjects were making choices according to a biased coin flip).

allowed us to easily and predictably manipulate subjects' beliefs about the proper world model. In particular, we manipulated belief in temporal dependence and coupling, factors that were important in generating a rational observer model for probability matching. To manipulate belief in temporal dependence, we varied whether outcomes would be interpreted as caused by factors external to the observer or as consequences of the observer's motor behavior. The rationale for this choice is that most external processes have temporal dependence, but outcome sequences generated by motor behavior are well modeled as independent (19). To manipulate belief in coupling, we provided cues to the similarity and ordering of the outcome probabilities, without affecting the actual outcomes.

In the first set of experiments, subjects played a series of "roulette games" (Fig. 2*A*). In the first condition ("even pieces, computer stops" condition), subjects viewed a "roulette wheel" with evenly sized colored slots. Subjects were instructed to choose one of the two colors, after which a black line rotated around the wheel before stopping on one of the two colors. A success was when the subject's choice matched the color upon which the line stopped. As in all standard sequential choice tasks, the colors had different associated probabilities (one color was correct on 67% of trials and the other on 33% of trials). In this condition, because the outcomes were generated externally (i.e., the computer program determined when and where the line would stop) and there was no good cue to coupling (i.e., the subjects were unable to correctly infer one option's success probability by simply looking at the other; if anything, because the pieces were evenly sized, it would promote the inference that they have equal probabilities of success), this environment was not suggestive of the true generative process and thus probability-matching behavior was predicted. In the second condition ("uneven pieces, subject stops" condition), two changes were made. First, the size of the colored pieces was no longer equal (one color's pieces were twice as large as the other). Second, subjects were led to believe that outcomes were dependent on their own motor behavior. To accomplish this, rather than having the computer determine exactly where the line would stop, subjects were instructed to press a key when the line was over a piece of their chosen color to "stop the line" (the actual outcomes were under complete experimental control; i.e., unbeknownst to the subjects the line could move one extra tick as necessary, thus ensuring the outcome-generating process was normatively identical across conditions) (*SI Appendix S3*).

Because subjects believed that success and failure were determined exclusively by their own motor skill, it should drastically change the world model they use to interpret the outcomes (see *SI Appendix*, Table S1 *A–C* for subject debriefing responses). In particular, motor behavior is one case where subjects are willing to posit temporally independent outcomes. Furthermore, whereas the task does not imply perfect coupling, it should foster the belief that what is learned from an attempt to stop on one color is at least partially informative about the probability of stopping on the other color (e.g., the probability of stopping on the big pieces should not be less than the probability stopping on the small pieces). Given these (correct) assumptions about the generative process for outcomes, the predicted behavior is therefore near maximizing.

As expected, probability-matching behavior was observed in the even pieces, computer stops condition, whereas near optimal behavior was observed in the uneven pieces, subject stops condition (Fig. 2*B*; average choice behavior significantly different in the two conditions, $t(11) = 2.9$, $P = 0.016$; even pieces, computer stops condition not significantly different from probability matching, 66%, $P = 0.75$; uneven pieces, subject stops condition significantly different from probability matching, 66%, $P < 0.001$; not significantly different from maximizing, 90%, $P = 0.37$—although we note that 90% is a somewhat arbitrary threshold). One possible alternative hypothesis is that subjects are actually probability matching in both conditions with the added supposition that subjects dramatically misestimate the probabilities of success in the uneven pieces, subject stops condition (i.e., they believe the probability of success for the better option is near 100%). Although this cannot be conclusively ruled out, it does appear to be inconsistent with subjects' reported beliefs (*SI Appendix S8*). Furthermore, the origins of such a belief would be unknown (and its persistence would be rather surprising) because it is inconsistent with the observed outcome statistics (to which subjects are quite sensitive; *SI Appendix S3*).

Multiple control conditions (*SI Appendix S4*, Fig. S1) demonstrated that optimizing behavior depended upon providing a model that was consistent with the true generative process (e.g., in conditions where subjects were also allowed to make a motor response if the pieces were evenly sized or if the probability of out-

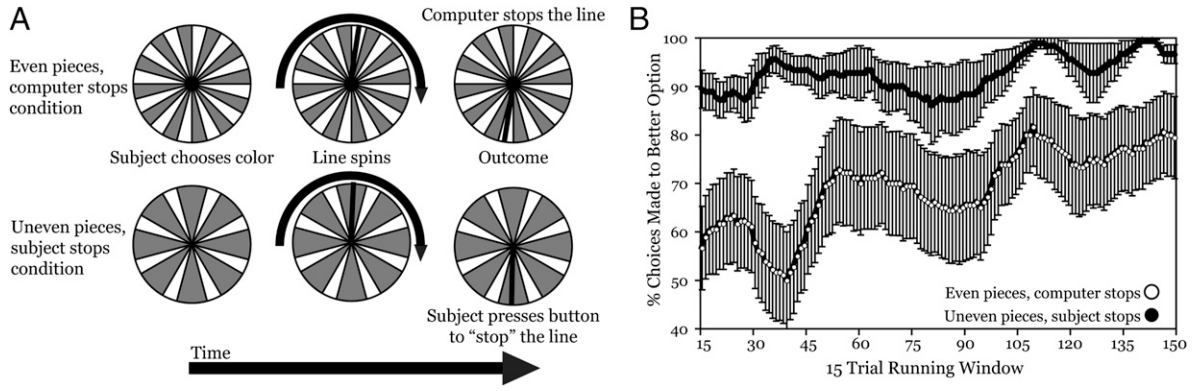

Fig. 2. Roulette experiment. (*A*) Experimental tasks. The basic visual stimulus was a roulette wheel made of pie pieces of alternating colors. In the even pieces, computer stops condition, the pieces were equally sized, thus giving no cue to the relative probability of success for the two options. In the uneven pieces, subject stops condition, the pieces for the high-probability color were twice as a large as the pieces for the low-probability color. In both conditions, the trial began with the subject selecting one of the two colors (here gray/white). A black line would then spin clockwise until, in the even pieces, computer stops condition, the computer stopped the line, or in the uneven pieces, subject stops condition, the subject stopped the line with a key press (the actual stop position being partially under experimental control). If the line stopped on a piece of the chosen color, this was considered a "success" and the subject won points. (*B*) Choice behavior. The percentage of choices to the better target in 15-trial moving windows is shown. In the even pieces, computer stops condition subjects show typical sequential binary choice behavior, starting near chance, and gradually reaching a plateau just above probability matching (i.e., probability matching+). Conversely, in the uneven pieces, subject stops condition, behavior is immediately maximizing. This difference in behavior can be attributed only to differences in subjects' beliefs about the generative process for the task, as the outcome statistics they observed were identical across conditions.

comes was reversed such that the bigger pieces were stopped on less frequently, then reasonably predictable suboptimal behavior emerged). Of particular note is the uneven pieces, computer stops control condition wherein, as in the uneven pieces, subject stops condition, subjects were given an explicit representation of the overall reward probabilities. However, despite this explicit representation, subject behavior was similar to that in the even pieces, computer stops condition, with significantly fewer choices toward the better option than in the uneven pieces, subject stops condition. This failure to maximize is consistent with previous work by Fantino and Esfandiari (4), which demonstrated suboptimal choice behavior even when subjects were explicitly told the overall reward probabilities, and also with the general fact that knowledge of the overall reward probabilities is typically insufficient to compute the optimal policy (i.e., it is only in the very special case where options are perfectly coupled and outcomes are completely independent across time that the overall reward probabilities are sufficient to compute the optimal policy).

Although the roulette task did effectively pit subjects' beliefs about the world model against the true generative process, we wished to demonstrate similar effects could be observed within the context of a more realistic and immersive environment and given more natural visuo-motor behaviors, in our case, aiming and shooting. Such an environment and task should elicit rich expectations about the world. In this series of experiments subjects piloted a "jet fighter" down a long tunnel (Fig. 3*A* and *Methods*) and at regular intervals they encountered pairs of colored targets. For each pair they were to select one target, aim their ship, and fire a bullet at the target. The target could either explode (success) or not (failure). In the first condition ("shield" condition), which was meant to be somewhat analogous to the standard binary choice task, when the subjects' bullet struck a target it either exploded or put up a previously invisible shield (with the better color rigged to explode on 70% of trials and the shield on 30% of trials). In a second condition ("different movement" condition), subjects were informed that if their bullet struck a target, it would explode (i.e., no shields). Instead, the targets were moved randomly ("jittered") in the tunnel, thus creating the illusion that motor skill determined whether a target was "hit" or "missed." Furthermore, one of the two targets jittered more rapidly than the other (and was also visually slightly smaller), which in the real world would tend to suggest it would be more difficult to hit than its slower, larger, partner. In reality, whether a target was hit or missed was determined experimentally via scripts attached to the target that either moved the target into the bullet (on trials where the target was the correct choice) or dodged the bullet (on trials where the target was the incorrect choice), thus ensuring that the conditions were in fact normatively identical (*SI Appendix S6*).

As in the roulette tasks above, the jet-fighter conditions were designed to be suggestive of either the wrong world model (externally generated outcomes = temporal dependence + poor/incorrect cue to coupling) or the correct world model (outcomes generated by motor behavior = temporal independence + cue to proper coupling) and, as was seen in the roulette tasks, these manipulations led to clear differences in behavior (Fig. 3*B*; average choice behavior significantly different in the two conditions, $t(11) = 3.2$, $P = 0.009$; shield condition not significantly different from probability matching, 70%, $P = 0.7$; different movement condition significantly different from probability matching, 70%, $P = 0.005$; not significantly different from maximizing, 90%, $P = 0.1$). Discrepancies in behavior from the roulette conditions may be due to the number of task-irrelevant factors inherent to the jet-fighter game that may nonetheless have affected behavior (i.e., the jitter could lead one target to be closer to the jet fighter than the other, thus affecting the value computation). Control conditions verified that the improvement in choice behavior required that the implied world model match the true generative process and was not a simple consequence of the targets moving and the shield being removed (e.g., if the targets appeared to move identically, or if the targets moved differently, but the success probabilities were reversed such that the "fast" target was actually "easier to hit," behavior was consistently suboptimal and, in fact, matched what would be expected given these erroneous world models; *SI Appendix S6* and *S7*, Fig. S2, and Table S2 *A* and *B*).

Discussion

Human sequential choice behavior depends on both the outcomes that are observed and beliefs about the process that generates those outcomes. When a task (such as the binary choice task as it has been traditionally implemented) provides poor cues to the true nature of the generative process, an erroneous model may be posited, which can lead to suboptimal and apparently irrational behavior. Other well-known irrational behaviors—such as the hot hand and gambler's fallacies (20)—also fit this pattern. These behaviors have often been explained via the "representativeness heuristic" (21), which can be qualitatively described as an incorrect

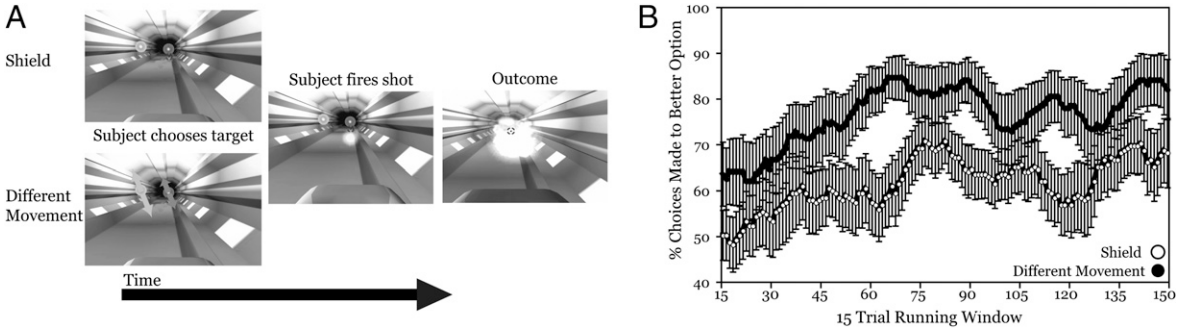

Fig. 3. Jet-fighter experiment. (*A*) Experimental tasks. In both the shield and the different movement conditions, subjects flew in a jet fighter down an endless tunnel and past pairs of colored targets. In the shield condition the targets were stationary with respect to the tunnel, thus giving no cue to the relative probability of success for the two options. In the different movement condition, the targets jittered relative to the tunnel, with the high-probability target jittering a smaller amount relative to the low-probability target. In both conditions, the subject was to choose one of the two targets, align their jet fighter with the target, and fire a bullet. In the shield condition, when the bullet struck a target it either exploded (success) or a previously invisible shield was deployed (failure). In the different movement condition, the bullet either struck the target (success) or missed (failure—again, although subjects believed success or failure was a function of motor skill, it was actually experimentally determined). (*B*) Choice behavior. The percentage of choices to the better target in 15-trial moving windows—first 150 trials. In the shield condition subjects show typical sequential binary choice behavior, starting near chance and gradually reaching a plateau just above probability matching (i.e., probability matching+). Conversely, in the different movement condition, choice behavior is immediately closer to optimal (although it does not reach optimal). This difference in behavior can be attributed only to differences in subjects' beliefs about the generative process for the task, as the outcome statistics they observed were identical across conditions.

belief about generative processes and, in particular, the temporal dependence of outcomes. Interestingly, individuals who demonstrate a propensity toward gambler's fallacy also show an inclination toward hot hand beliefs, suggesting that a strong belief in temporal dependence can manifest itself as a belief in either state transience or state persistence models depending on the circumstances (22). Conversely, when a task provides good cues to the true generative process, the advantages of a model-based system are immediately apparent. For instance, such a system allows one to make accurate predictions given relatively small amounts of actual outcome data. This fact is clearly observed in Fig. 2*B*, where subjects are able to determine the optimal policy almost immediately —certainly well before sufficient outcome data have accrued to warrant such a policy. Although learning a generative model from scratch is computationally taxing (which is one of the primary reasons model-free algorithms have risen to prominence in artificial intelligence/robotics applications), humans have a vast amount of prior knowledge that allows the flexible construction of plausible models. To illustrate, subjects did not need to "learn" that outcomes were generated independently; they could combine their belief that outcomes were generated via motor behavior with their knowledge that motor behavior is reasonably independent across time to reach the proper conclusion. Given the vast amount of prior knowledge acquired by every adult brain, it is sensible for humans to always be model-based decision makers (and thus individuals studying human decision making need to not only be aware of the data they make available to the subject, but also probe subjects' beliefs about the process generating the data).

Model-free reinforcement learning agents bypass the generative process for outcomes and instead derive policies from the observed rewards (23). Thus, such an agent could not produce the observed pattern of results (because the outcome statistics were identical across conditions and thus the agent's behavior would be identical across conditions as well). This evidence for model-based decision making poses a challenge for the most popular theories regarding the neural bases of human choice behavior, which have focused on the correspondence between activity in various anatomical areas, such as the ventral striatum, and values, such as prediction errors, calculated via model-free algorithms (18, 24, 25). It has been suggested (26) that two different systems are involved in choice behavior, one that computes a model-free estimate of value (striatum) and another that is involved in model-based/predictive computations (prefrontal cortex). Although it is certainly possible that depending on the exact task subjects will behave more like model-free than model-based learners, our data suggest that it is actually the latter, rather than the former, that is primarily driving choice behavior in the sequential binary choice task.

Methods

Bayesian Model (See *SI Appendix S1* for More Details). Most experiments involving two-option sequential choice tasks generate a sequence of independently sampled binary outcomes x from a Bernoulli distribution—i.e., binary outcomes (success/failure) with fixed outcome probabilities (p and $1-p$). A learning agent that does not know this generative process has uncertainty about both the coupling and the temporal dependence of outcomes. To illustrate, consider an example card task. On each trial the agent chooses one of the two cards and either wins money for selecting the right option (success) or wins nothing (failure). For a learning agent, the observable signal is the payoff. Whereas the experimenter knows that the options are coupled, such that when one yields a payoff the other yields nothing, a learning agent cannot be sure what would have happened if the other option had been selected. In addition, without prior knowledge, a learning agent cannot know that outcomes are temporally independent. The learning agent needs to represent all these possibilities. Moreover, to make intelligent choices, it must be able to predict the reward probabilities for the next trial.

The base of the model is a one-step transition matrix that encodes the probability that an option will have a certain outcome given the outcome that was observed on the previous trial. For instance, assume that the agent selected option 1 and the observed outcome was a success. The next decision is affected by the probability that option 1 switches to failure on the next trial given that it was a success on the current trial—$P(x^1_{t+1}=0 \mid x^1_t=1)$—and the probability that option 2 becomes a success, given that it could have been a success or a failure on the current trial (because option 2 was not selected and thus no outcome was observed)—$P(x^2_{t+1}=1 \mid x^2_t=1)$ or $P(x^2_{t+1}=1 \mid x^2_t=0)$. The model keeps track of all possible next outcomes given any current observation. There are four possible hypotheses about the next outcome of the two options (neither option is a success, only option 1 is a success, only option 2 is a success, and both options are a success). These possibilities can be written in matrix form, given the four possible states of the current options,

$$P\left(x^1_t \mid x^1_{t-1}\right) = \begin{bmatrix} 1-\alpha^1_{0\to1} & \alpha^1_{1\to0} \\ \alpha^1_{0\to1} & 1-\alpha^1_{1\to0} \end{bmatrix}$$

$$P\left(x^2_t \mid x^2_{t-1}\right) = \begin{bmatrix} 1-\alpha^2_{0\to1} & \alpha^2_{1\to0} \\ \alpha^2_{0\to1} & 1-\alpha^2_{1\to0} \end{bmatrix},$$

where $\alpha^i_{0\to1}$ is shorthand notation for $P(x^i_{t+1}=1 \mid x^i_t=0)$: the probability of switching from failure at time t to success at time $t+1$, and similarly $\alpha^i_{1\to0}=P(x^i_{t+1}=0 \mid x^i_t=1)$. In this form, different possible structural assumptions about the environment can be expressed as dependencies between the parameters. For example, if trials are independent, then $\alpha^i_{0\to1}=1-\alpha^i_{1\to0}$—the probability of an option transitioning from failure to success ($\alpha^i_{0\to1}$) is the same as that from success to success ($\alpha^i_{1\to1}$, which is equal to $1-\alpha^i_{1\to0}$). In other words, if trials are independent, then the probability of success does not depend on what occurred on the previous trial. If options are perfectly coupled, the probability of transitioning from success to failure for one option is 1 – the other: $\alpha^2_{0\to1}=1-\alpha^1_{0\to1}$ and $\alpha^2_{1\to0}=1-\alpha^1_{1\to0}$. In other words, on every trial, one option is a success and one is a failure and thus if one transitions, the other has to as well.

We adopt a Bayesian framework where the agent learns the value of these parameters, and model assumptions are built into priors on these parameters (*SI Appendix*). In the model presented in Fig. 1, the model was initialized with the belief that outcomes were "sticky" (i.e., successes tend to transition to successes and failures tend to transition to failures) and that the options were uncoupled (i.e., observing an outcome after selecting option 1 says nothing about what would have occurred had option 2 been selected). Bayesian learning in the model is extremely simple; it amounts to incrementing the appropriate parameter after each observed transition. However, which transitions are observed depends on the way actions are selected. We assume that action selection is based on a one-step look-ahead prediction of the reward state for both options, where actions are selected deterministically on the basis of the option with the highest expected value (i.e., a max decision rule).

The data presented in Fig. 1 are the result of 10 such agents performing the same sequential task that is described in *SI Appendix S2* [450 trials, p (better option) = 0.6].

Roulette Decision Task (See *SI Appendix S3* for Outcome Statistics). ***Subjects.*** Twelve subjects (four males, mean age = 20.1 y) participated in both conditions described below (as well as an additional condition described in *SI Appendix S4*). Run order of the task conditions was counterbalanced. Subjects provided written informed consent to participate and were paid $10/h.

Apparatus. The apparatus consisted of a Dell XPS running Windows XP and MATLAB (Math Works) and the Psychophysical Toolbox (http://psychtoolbox.org) (27, 28). The stimuli were displayed on a 20-in Dell LCD monitor at a resolution of 1,680 × 1,050 pixels by a NVIDIA GeFORCE 8800 GTX video card. Subjects were seated ≈59 cm from the screen.

Stimuli/conditions. **Roulette condition 1: Even pieces, computer stops condition.** The roulette wheel consisted of a large circle (15° diameter) divided into individual colored wedges (3° each for a total of 120 wedges). On each trial (150 total trials) the subject was instructed to choose one of the two possible colors (orange or purple) after which a solid black line rotated around the circle (180°/s) before stopping on one of the two colors (with the probability of landing on the high-probability color being 67%). If the chosen color matched the color on which the line stopped, the outcome was considered a success and, if not, it was considered a failure. Subjects received auditory feedback as well as +10 points on successful trials, with their instructions being to gain as many points as possible.

Roulette condition 2: Uneven pieces, subject stops condition. The task was the same as above with two changes. First, the size of the colored wedges was no longer equal. The wedges for the low success probability wedges remained 3°; however, the high-probability wedges were increased in size to 6°. Second, subjects were instructed that after they made their color choice and the

line began moving, they were to press the spacebar when the line was over a piece of their chosen color to "stop" the line on that piece. Whereas subject motor behavior did determine the approximate area in which the line stopped, the specific wedge was determined experimentally such that in 67% of trials the line would stop on a high-probability wedge (if necessary, the line would move one additional tick beyond where the subject actually responded—this deception was not noted by any subject).

Jet-Fighter Decision Task (See Also *SI Appendix S5, S6,* and *S7*). ***Subjects.*** Thirteen subjects participated in both conditions described below (as well as an additional condition described in *SI Appendix S7*). One female subject was removed before data analysis for not following instructions (did not attempt to gain as many points as possible). The final group therefore consisted of 12 subjects (three males, mean age = 21.3 y). Run order of the task conditions was counterbalanced. Subjects provided written informed consent to participate and were paid $10/h.

Apparatus. The apparatus consisted of a Dell XPS running Windows XP and Virtools Dev 3.5.0.24, which was used to display stimuli and collect the data. The stimuli were displayed on a 50-inch Panasonic HD plasma television model TH-50PZ700U driven at a resolution of 1,680 × 1,050 pixels by a NVIDIA GeFORCE 8800 GTX video card. Subjects were seated 1.7 m from the screen.

Stimuli/conditions. **General jet-fighter environment.** In each of the following experiments subjects flew a simulated jet fighter through an endless tunnel. For ease of exposition, all units are given in units native to the game environment [Virtools units (vu); at a simulated depth of 0 cm, 1 vu ≈0.4 cm]. Subjects could move the jet fighter from side to side and up and down (*x* and *y* directions, respectively) using the arrow keys on the keyboard, but their speed through the tunnel (*z* direction) was kept at a constant of 900 vu/s. As the subjects flew through the tunnel, they encountered pairs of colored target spheres, 30 vu in size. The subjects' task was to align the ship with one of the two targets and once properly aligned and within range, they were to fire a "bullet" at the target using the spacebar (only one bullet could be fired per target pair). The outcome of this action depended on both their alignment accuracy and the task conditions (see below). The next target pair was loaded a simulated depth of 1,350 vu at positions equidistant from the position of the jet fighter at load time (ensuring that relative proximity should not be a factor in the subjects' decisions) and became visible after the ship passed the plane of the previous target pair. Subjects therefore encountered a target pair approximately once every 1.5 s.

Experimental task conditions. Before the experimental tasks, subjects were trained to navigate and shoot in the environment (*SI Appendix S5*). A total of 450 trials were completed in shield condition. In the different movement condition, 600 trials were completed due to concerns that poor accuracy could leave too few trials to analyze if only 450 trials were completed. However, this fear proved unfounded and all analyses consider only the first 450 trial conditions. Subjects were given a short break at the halfway point of each condition.

Jet-fighter condition 1: Shield condition. In the stationary target condition, on each trial one of the two targets (better target, orange; worse target, sky blue) was selected to "explode" upon a hit by a bullet and the other was selected to shield upon a hit. The hit outcome was determined by a biased coin flip such that in 70% of trials the better target was selected to explode upon a hit by the players' bullet, and in the remaining 30% of trials the worse target was selected to explode.

Jet-fighter condition 2: Different movement condition. The targets in this condition (better target, red; worse target, sea green) jittered randomly in the *x* and *y* directions with the constraints that they stay on their own side of the tunnel and avoid close contact with one another. In addition to color, the targets could also be differentiated by size and by movement characteristics. The visual size of the low success probability target was decreased by one-third as compared with the better target (20 vu vs. 30 vu). Note that because the explode radius remained the same, this change in visual size did not actually affect the ability of a subject to hit the target. Second, the degree of jitter was different for the two targets. Specifically, the low success probability target moved much faster with greater amplitude jumps (jitter drawn from a uniform distribution between −1.5 vu and 1.5 vu in both *x* and *y* directions on every frame) than the high success probability target (distribution between −1 vu and 1 vu).

Importantly, these visual differences were not actually causally related to the probability of hits and misses. On each trial, one of the two targets was selected to "intercept" a fired shot within the script activation zone (within 70 vu of the target center) and explode upon bullet impact, and the other was selected to "dodge" any shot fired within the script activation zone (with the two targets differing in the probability of an intercept/dodge as above—high-probability target = 70% intercept, 30% dodge). Shots fired outside of either target's script activation zone always led to a miss. The scripts were weaved into the random jittering in such a way that no subjects reported observing anything other than random movement from the targets.

ACKNOWLEDGMENTS. This research was supported by the Office of Naval Research Grant N00014-07-1-0937.

1. Gaissmaier W, Schooler LJ (2008) The smart potential behind probability matching. *Cognition* 109:416–422.
2. Stanovich KE, Stanovich KE (2003) Is probability matching smart? Associations between probabilistic choices and cognitive ability. *Mem Cognit* 31:243–251.
3. Unturbe J, Corominas J (2007) Probability matching involves rule-generating ability: A neuropsychological mechanism dealing with probabilities. *Neuropsychology* 21:621–630.
4. Fantino E, Esfandiari A (2002) Probability matching: Encouraging optimal responding in humans. *Can J Exp Psychol* 56:58–63.
5. Vulkan N (2000) An economist's perspective on probability matching. *J Econ Surv* 14: 101–118.
6. Fiorina MP (1971) A note on probability matching and rational choice. *Behav Sci* 16: 158–166.
7. Gal I, Baron J (1996) Understanding repeated simple choices. *Think Reason* 2:81–98.
8. Shanks DR, Tunney RJ, McCarthy JD (2002) A re-examination of probability matching and rational choice. *J Behav Decis Mak* 15:233–250.
9. Rode C, Cosmides L, Hell W, Tooby J (1999) When and why do people avoid unknown probabilities in decisions under uncertainty? Testing some predictions from optimal foraging theory. *Cognition* 72:269–304.
10. Wolford G, Newman SE, Miller MB, Wig GS (2004) Searching for patterns in random sequences. *Can J Exp Psychol* 58:221–228.
11. Rubinstein I (1959) Some factors in probability matching. *J Exp Psychol* 57:413–416.
12. Hake HW, Hyman R (1953) Perception of the statistical structure of a random series of binary symbols. *J Exp Psychol* 45:64–74.
13. Goodie AS, Fantino E (1999) What does and does not alleviate base-rate neglect under direct experience. *J Behav Decis Mak* 12:307–335.
14. Goodnow JJ (1955) Determinants of choice-distribution in two-choice situations. *Am J Psychol* 68:106–116.
15. Yellot JI (1969) Probability learning with noncontingent success. *J Math Psychol* 6: 541–575.
16. Feldman J (1959) On the negative recency hypothesis in the prediction of a series of binary symbols. *Am J Psychol* 72:597–599.
17. Edwards W (1961) Probability learning in 1000 trials. *J Exp Psychol* 62:385–394.
18. Schönberg T, Daw ND, Joel D, O'Doherty JP (2007) Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making. *J Neurosci* 27:12860–12867.
19. Faisal AA, Selen LPJ, Wolpert DM (2008) Noise in the nervous system. *Nat Rev Neurosci* 9:292–303.
20. Lee W (1971) *Decision Theory and Human Behavior* (Wiley, New York).
21. Kahneman D, Tversky A (1972) Subjective probability: A judgment of representativeness. *Cognit Psychol* 3:430–454.
22. Sundali J, Croson R (2006) Biases in casino betting: The hot hand and the gambler's fallacy. *Judgm Decis Mak* 1:1–12.
23. Sutton RS, Barto AG (1998) *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, MA).
24. Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ (2006) Cortical substrates for exploratory decisions in humans. *Nature* 441:876–879.
25. Wittmann BC, Daw ND, Seymour B, Dolan RJ (2008) Striatal activity underlies novelty-based choice in humans. *Neuron* 58:967–973.
26. Daw ND, Niv Y, Dayan P (2005) Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat Neurosci* 8:1704–1711.
27. Brainard DH (1997) The Psychophysics Toolbox. *Spat Vis* 10:433–436.
28. Pelli DG (1997) The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spat Vis* 10:437–442.