HISTORICAL NEWS & VIEWS: DECISION NEUROSCIENCE

# Are we of two minds?

What you choose depends on what information your brain considers and what it neglects when computing the value of actions. An early theory used this insight for a computational account of habits versus deliberation. It has ultimately helped uncover how choice in the brain goes beyond such simple dichotomies.

Nathaniel D. Daw

The idea that our decisions arise from multiple competing systems—cold versus hot, deliberative versus automatic—is ubiquitous, underpinning folk and scientific intuition about conflict, temptation, and self-control. Yet it is also deeply puzzling. Why evolve two choice mechanisms? After all, this doesn't actually solve the problem of deciding. To the contrary, it compounds it, since your brain must also decide which system to trust. Plato analogized this problem to that of a charioteer wrangling a pair of horses, one noble and one beastly. But how does the charioteer work, or for that matter the horses? And why not just ride the noble one?

In an early theoretical article in *Nature Neuroscience*[1], we looked to machine learning for answers to these questions. Ideally, one would choose the action with the largest expected utility, which doesn't seem to leave room for multiplicity. But in realistic tasks, like spatial navigation—in which every choice leads to more choices, such that their consequences are delayed and contingent—even computing this expectation is laborious or unfeasible, since it requires enumerating all possible paths. Algorithms in reinforcement learning (solving choice problems by trial and error) must employ a range of shortcuts to approximate this ideal. A key distinction is that between model-based (MB) and model-free (MF) algorithms. MB algorithms learn a representation ('internal model') of the task contingencies—like a map—which they use to compute the expected value of candidate actions by iteratively tracing out their consequences (Fig. 1a). Though accurate, such simulation is laborious, so MF algorithms avoid it by instead storing the endpoint of all this computation: the long-run expected value of each action. This simplifies choice, at the cost of inflexibility: if the world changes, the stored values may be invalid and produce outdated choices. For instance, actions leading to food should carry less value once I am full.

We proposed[1] that the MB-versus-MF distinction formalized a well-supported dual-system theory from behavioral neuroscience: the distinction between goal-directed and habitual instrumental behavior[2]. A form of MF learning known as temporal-difference learning was already the predominant theory of the midbrain dopamine system and its role in reinforcing successful actions in striatum. We pointed out that this corresponded well to inflexible, habitual behaviors that arise after overtraining[2]. However, that mechanism failed to explain how animals can also flexibly solve decision problems that seemed to require a world model. By proposing a MB system alongside the MF one, we brought mental simulation and goal-directed choice into the same computational framework.

One key problem for testing multiple decision system theories in the laboratory is that any reasonable decision system will try to maximize reward, and so, they will often all make similar choices. Hence it is often ambiguous which hypothetical system is responsible for a particular behavior and how to interpret any effects on choice of (for instance) neural manipulations (was the beastly or the noble horse affected?). Historically, this has motivated laborious experimental procedures to contrive circumstances in which the systems' contributions can be differentiated. For instance, habitual behaviors persist following reward devaluation. Formalizing the putative systems in terms of MB and MF algorithms offered a nimbler approach to this problem, because they make concrete, distinct predictions about how subjects will adjust their choices in light of each trial's outcome in multistep decision tasks[3].

Relative to earlier, one-shot manipulations like reward devaluation, this learning-based approach to differentiating the systems is better suited to dynamic neuroscientific measurement. In human neuroimaging, correlates of trial-by-trial MB and MF decision variables have now been observed throughout a broad network[3,4]. One theme of these studies is that these systems appear to be less separated in the intact brain than might have been expected based on earlier animal lesion studies.

People's learning behavior on such tasks has also been used extensively to study the tradeoff between MB and MF processes in humans, and in particular, to document circumstantial and individual differences affecting the degree to which people rely on internal models. Part of the appeal of dual-system theories is that they offer an intuitive explanation for why people might be of two minds about something[5], an idea that has been invoked in such diverse areas as moral dilemmas, racism, and self-control. Having a formal definition of the systems has helped to investigate these claims experimentally. For instance, a range of psychiatric symptoms involving compulsion (from morning drinking to repetitive hand-washing) is associated with reduced MB behavior on a reinforcement learning task[6]. This supports the longstanding suggestion that imbalance in deliberative versus automatic processes might underlie the compulsive character of disorders such as drug abuse[5].

A further headline claim of the original theory[1] was an account of the charioteer problem: how the brain might arbitrate between the systems. The idea was that different algorithms are most trustworthy in different circumstances, so action evaluations should compete on the basis of their statistical reliability, or conversely, their uncertainty. This offered a rational take on automaticity and control. Later theories refined the account to speak more explicitly of the tradeoff between the costs (time) versus benefits (better choices, harvesting more rewards) of MB computation, relative to quickly executing a MF habit[7,8]. Thus, for instance, when learning a new uncertain task, it might behoove you to deliberate about the consequences of your actions (MB), but after lengthy practice on a stable task, you can usually get the same result quicker by repeating what has always worked (MF). Such reasoning explains

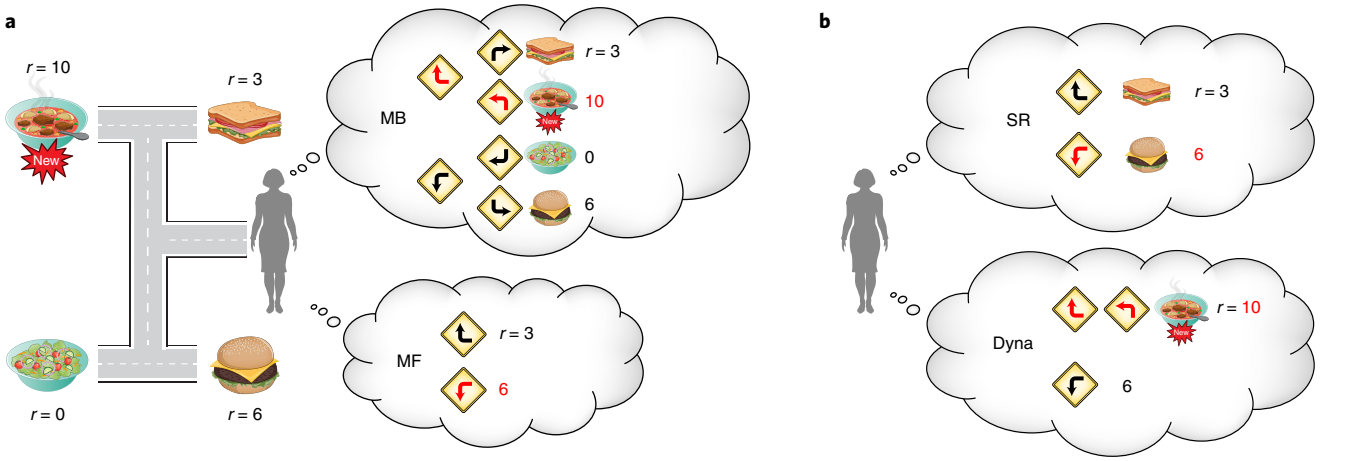


**Fig. 1 | Different evaluation strategies choose different routes to lunch. a**, MB and MF evaluation. Different sequences of actions lead to different food outcomes. The agent had previously favored a left turn for a burger with a reward value $r = 6$. She has now learned of, but not yet traveled to, a new hearty soup restaurant, whose higher value ($r = 10$) has been encoded in her internal model. MB evaluation, which computes actions' values by exhaustive mental simulation, thus prefers to turn right at the start to obtain the soup (red). An MF system instead relies on a summary of the previously experienced long-run value of each action. This reliance on stored summaries can lead to inappropriate choices if they become out of date: because the MF value is learned from experiencing an action's consequences, it does not yet reflect the soup's availability and still prefers a left turn. **b**, Evaluation strategies intermediate between MB and MF. The successor representation (SR) stores a simplified model, summarizing the long-run experienced outcome of the action (for example, the hamburger). Compared to MF, this allows for greater flexibility when an outcome's value changes, but in the current example, like MF, this summary neglects the availability of the new path to soup, and thus still prefers the hamburger. Finally, hybrid algorithms like Dyna maintain a set of MF values, but selectively update them using individual model evaluations that can be prioritized as needed. Here information about novel soup can trigger recomputation of the value leading there.

the emergence of habits with overtraining. There is still much experimental work to be done to test such ideas of dynamic arbitration—even to clarify what is the space of potential competing models—though initial forays are promising[4].

Of course, such clean dichotomies are bound to be oversimplified. In formalizing them, the MB-versus-MF distinction has also offered a firmer foundation for what will ultimately be, in a way, its own undoing: getting beyond the binary. Neither MB recomputation nor MF reuse need be complete, and researchers are beginning to find evidence that brains combine these approaches to produce various intermediate strategies. The key insight remains that the phenomena associated with dual-system conflict (habits, slips of action) reflect the brain adaptively deploying and reusing decision computations, even if this is not simply by switching between discrete systems.

For instance, the basic MF trick of storing computations can be applied more judiciously. Rather than reusing completed evaluations, it is possible to store intermediate steps of value computation, such as expectancies about long-run action outcomes (for example, if you head in the direction of the drive-through, you will ultimately get a hamburger). Such an approach, known as the successor representation[9] (Fig. 1b), permits finalizing plans relatively quickly (you need only to combine the expectation of a hamburger with its current value, without considering the intervening steps). This is more flexible than MF choice at coping with change in outcome values (for example, an *E. coli* outbreak). But reusing old predictions can still cause mistakes when other aspects of the environment change (for example, if the drive-through closes, or a new hearty soup restaurant opens nearby). Such fingerprints of this valuation strategy have also been reported in people[10], offering an example subtler than outright habits of how selective computation can misfire.

Conversely, MB simulation can't possibly anticipate all possible future paths in most environments. Accordingly, it must truncate its simulations and focus them in particular directions[11]. This observation suggests that rather than selection between exhaustive MB computation and none at all, control of evaluation is better understood as selection over which particular paths to consider and when[8] (as in algorithms like Dyna[12]; Fig. 1b). This more granular perspective may begin to extend the explanatory power of the theory beyond outright neglect—as in habits, compulsive actions, and the like—and help to explain more directed phenomena, like rumination, craving, and the effects of advertising.

Newer theories that better detail the individual steps of MB computation should also help to guide research into a key remaining question in the area: what neural mechanisms carry out these computations? We've long had a basic picture of how dopaminergic prediction errors could support MF learning. Although the original theory[1] envisioned that this canonical habit circuit would compete against a separate MB system, evidence since then has instead suggested that MB computation, too, shares a dopaminergic foundation[13]. The intermediate computational strategies discussed above suggest how this might work, with MB predictions selectively layered over a shared MF learning stage[8].

Beyond these coarse, systems-level suggestions, the choice-time mechanisms by which the brain computes MB valuations remain open for discovery. Although experimental work on these ideas has so far primarily centered on studies in human subjects (where methods like online testing permit rapidly refining tasks and analyses), this question ultimately calls for measuring and manipulating neural events with finer temporal resolution. Accordingly, researchers have begun to adapt approaches from humans to nonhuman animals, where they can be combined with more invasive methods[14]. One of the most promising frontiers is in the hippocampus, where

researchers have observed activity tracking nonlocal paths (for example, ahead or behind the animal) in the same neurons that usually represent the animal's current location[15]. This may be a direct window into individual trajectories of MB 'mental simulation', and indeed many of the regularities of these nonlocal trajectories are explained by the hypothesis that they are adaptively selected to optimize planning[8].

In the end, perhaps we are not creatures of two minds—or three, or four—but it has become increasingly clear that what we choose depends to a surprising extent on how we compute the values of our candidate actions. And there are many different, interacting routes to this evaluation. ❐

Nathaniel D. Daw

*Princeton Neuroscience Institute and Department of Psychology, Princeton University, Princeton, NJ, USA.*
*e-mail: ndaw@princeton.edu*

Published online: 22 October 2018
https://doi.org/10.1038/s41593-018-0258-2

## References

1. Daw, N. D., Niv, Y. & Dayan, P. *Nat. Neurosci.* **8**, 1704–1711 (2005).
2. Balleine, B. W. & Dickinson, A. *Neuropharmacology* **37**, 407–419 (1998).
3. Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P. & Dolan, R. J. *Neuron* **69**, 1204–1215 (2011).
4. Lee, S. W., Shimojo, S. & O'Doherty, J. P. *Neuron* **81**, 687–699 (2014).
5. Everitt, B. J. & Robbins, T. W. *Nat. Neurosci.* **8**, 1481–1489 (2005).
6. Gillan, C. M., Kosinski, M., Whelan, R., Phelps, E. A. & Daw, N. D. *eLife* **5**, e11305 (2016).
7. Keramati, M., Dezfouli, A. & Piray, P. *PLoS Comput. Biol.* **7**, e1002055 (2011).
8. Mattar, M. G. & Daw, N. D. *Nat. Neurosci.* https://doi.org/10.1038/s41593-018-0232-z (2018).
9. Dayan, P. *Neural Comput.* **5**, 613–624 (1993).
10. Momennejad, I. et al. *Nat. Hum. Behav.* **1**, 680–692 (2017).
11. Huys, Q. J. et al. *PLoS Comput. Biol.* **8**, e1002410 (2012).
12. Sutton, R. S. Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. *Proc. Int. Conf. Mach. Learn.* **7**, 216–224 (1990).
13. Sharpe, M. J. et al. *Nat. Neurosci.* **20**, 735–742 (2017).
14. Miller, K. J., Botvinick, M. M. & Brody, C. D. *Nat. Neurosci.* **20**, 1269–1276 (2017).
15. Foster, D. J. *Annu. Rev. Neurosci.* **40**, 581–602 (2017).

## Acknowledgements

I am grateful to my coauthors on the original article, Y. Niv and P. Dayan, and to many other collaborators, trainees, and colleagues who have helped to develop these ideas over the intervening years.

## Competing interests

The author declares no competing interests.