The Role of Orbitofrontal Cortex in Decision Making

A Component Process Account

LESLEY K. FELLOWS

Montreal Neurological Institute, McGill University, Montréal, Québec, Canada

ABSTRACT: Clinical accounts of the effects of damage to orbitofrontal cortex (OFC) have provided important clues about the functions of this region in humans. Patients with OFC injury can demonstrate relatively isolated difficulties with decision making, and the development of laboratory tasks that captured these difficulties was an important advance. However, much of the work to date has been limited by the use of a single, complex decision-making task and by a narrow focus on risky decisions. A fuller understanding of the neural basis of decision making requires identification of the simpler components that underlie this complex behavior. Here, I review evidence that OFC lesions disrupt reversal learning in humans, as in animals, and show that this deficit in reversal learning is an important mechanism underlying the difficulties of such patients in the Iowa gambling task. Reversal learning, in turn, can be decomposed into simpler processes: a failure to rapidly learn from negative feedback may be the critical difficulty for OFC patients. OFC damage can also affect forms of decision making that do not require trial-by-trial learning. Preference judgment is a simple form of decision making that requires comparing the relative value of options. Humans with OFC lesions are more inconsistent in their choices, even in very simple preference judgment tasks. These results are broadly consistent with the view that OFC is critically involved in representing the relative value of stimuli, but also raise the possibility that this region plays distinct roles in reinforcement learning and value-based judgment.

KEYWORDS: reversal learning; neuroeconomics; executive function; prefrontal cortex; human; lesion

Clinical descriptions of patients with ventral frontal lobe damage have heavily influenced current thinking about the functions of orbitofrontal cortex (OFC) in humans. However, despite vivid anecdotal accounts of social, emotional, and personality changes following OFC injury stretching back many

Address for correspondence: Lesley K. Fellows, M.D., C.M., DPhil, Montreal Neurological Institute, 3801 University St., Rm 276, Montréal, QC H3A 2B4, Canada. Voice: (514) 398 8980; fax: (514) 398 1338.
lesley.fellows@mcgill.ca

Ann. N.Y. Acad. Sci. 1121: 421–430 (2007).
doi: 10.1196/annals.1401.023

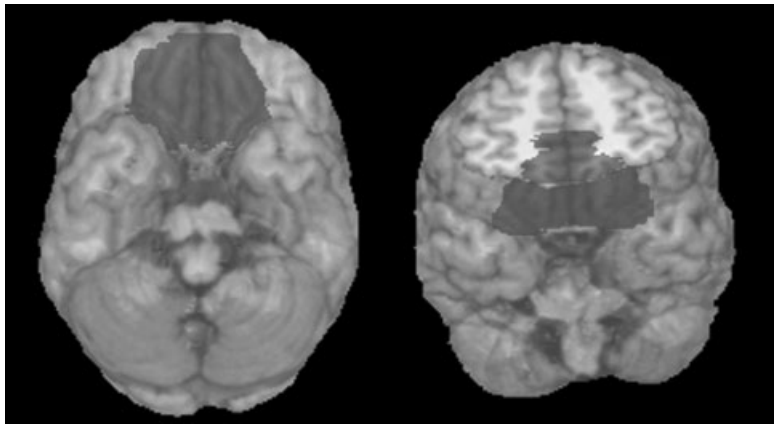

FIGURE 1. Schematic representation of the region of the frontal lobes referred to as ventromedial frontal (VMF) in the text. This includes medial OFC (shown in dark grey in the base of the brain view in the left panel) and the adjacent ventral region of medial PFC (shown in oblique view, with the anterior portion of the frontal lobe cut away, in the right panel). The common causes of focal injury to these areas in humans typically affect both sectors, often bilaterally, albeit to varying degree.

decades,[1–4] a principled understanding of the basis of these deficits has been slow to emerge. As with many disorders of complex behavior, the sticking point has been how best to frame these clinical observations: One influential model proposes that a fundamental impairment in decision making is at the heart of the real-life difficulties of OFC-damaged patients.[5] Experimental evidence for this claim came from the observation that those patients with OFC damage who displayed clinical evidence of impaired decision making were also impaired on a laboratory decision task now known as the Iowa gambling task (IGT).[6] Prompted in part by these findings, other investigators began asking more general questions about how economic information important to decision making, such as expectancies, risk, and uncertainty, might be represented in the brain.[7–10] This line of research, sometimes called "neuroeconomics," has provided evidence that activity within OFC (and the anatomically closely related ventral aspect of medial prefrontal cortex [PFC]; FIG. 1) reflects the relative value of potential choices (see Padoa-Schioppa *et al.*—this volume).

Although it was developed to study decision making, the IGT could equally be viewed as a reinforcement learning task: Good performance requires learning the reward and punishment contingencies associated with the different decks of cards and integrating these varying contingencies over multiple trials. Could the poor performance of OFC-damaged patients reflect a basic difficulty in some aspect of reinforcement learning? This formulation of the problem brings to bear a different literature. There is abundant evidence for a role for OFC in specific forms of reinforcement learning, primarily from

animal studies. In particular, OFC lesions in several species lead to a characteristic deficit in reversal learning (reviewed in Ref. 11 and elsewhere in this volume). Two features of the initial IGT work in patients with frontal damage raise the possibility that reversal learning might be playing a role in this task. First, the task itself involves a reversal of initial reward and punishment contingencies. Second, at least in the initial IGT study, patients with ventromedial frontal (VMF) damage failed to learn to avoid the disadvantageous decks,[6,12] a behavior that echoes the tendency of animals with OFC damage to perseverate on the initially rewarded stimulus after reinforcement contingencies change in reversal-learning tasks.

The IGT requires participants to choose among four decks of cards. On each trial, a card is drawn, which either provides a win, or a win and a loss. Overall, two of the decks are associated with large wins, but even larger losses. The other two ('advantageous') decks provide smaller wins, but even smaller losses. Crucially, the order of the cards in each deck is fixed. The large losses associated with the disadvantageous decks only begin to accrue after several trials in which only large wins are experienced.[13] Unsurprisingly, healthy controls and patients alike show a preference for these (eventually) disadvantageous decks in the first block of 20 trials, because the reinforcement contingencies that have been experienced up to that point indicate that these decks are "the best bet." As the task proceeds, and the large losses begin to accrue, healthy subjects gradually shift their choices to the two advantageous decks. In contrast, those with VMF damage persist in choosing more often from the initially attractive, but overall disadvantageous decks. We hypothesized that the specific pattern of reinforcement in the IGT required reversal learning, and that the persistently disadvantageous choices of VMF patients reflected a fundamental impairment in reversal learning similar to that seen in other species after OFC damage.

This hypothesis was tested in work I carried out with Martha Farah. We first asked whether VMF damage in humans impaired simple reversal learning, following up an earlier lesion study which suggested as much.[14] Eight subjects with fixed focal damage to VMF due to stroke or aneurysm rupture were compared to 12 subjects with damage to other areas of the frontal lobes, and to 12 healthy, demographically matched control subjects. The reversal-learning task was a simple, two deck card game. Choosing from one deck led to a $50 play-money win, from the other a $50 loss. Once subjects had chosen from the winning deck eight trials in a row, the contingencies were switched without warning. The task continued for a further 50 trials, allowing up to five reversals. All subjects were quick to learn the initial associations in this simple task. However, those with VMF damage made substantially more errors in the reversal phase[15] (FIG. 2).

Having confirmed that VMF damage specifically impaired reversal learning, we went on to ask whether this explained the characteristic difficulties with the IGT experienced by patients with VMF damage. To that end, we designed

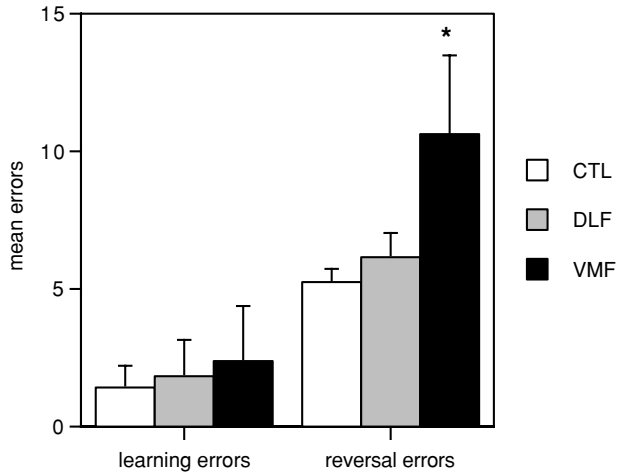

FIGURE 2. Initial stimulus−reinforcer association learning and reversal-learning performance in subjects with fixed damage of VMF lobes (VMF), compared to subjects with damage to the frontal lobes outside VMF (DLF) and healthy control subjects (CTL). Initial learning performance is expressed as the mean number of errors made before the learning criterion was met and reversal learning as the mean number of errors in the reversal phase of the experiment. Error bars show the upper bound of the 95% confidence intervals. The asterisk indicates a significant group X error type interaction, $P < 0.05$. From Fellows, L.K. & M.J. Farah. Ventromedial frontal cortex mediates affective shifting in humans: evidence from a reversal learning paradigm. *Brain* 2003; 126: 1830–1837, by permission of Oxford University Press.[15]

a "shuffled" variant of this task. By changing the order of the cards in each deck so that the large losses associated with the disadvantageous decks were experienced early, we attempted to eliminate the reversal-learning requirement of the original task. All other features of the task were identical to the original. Ten subjects with VMF damage (again compared to subjects with non-VMF frontal damage [N = 12] and healthy control subjects) completed both the original and "shuffled" versions of the IGT. As in previous reports, those with VMF damage chose from the disadvantageous decks more often than did the healthy control group. However, when the reversal requirement was eliminated in the shuffled task, the performance of VMF subjects was indistinguishable from that of healthy controls (FIG. 3).[16] Interestingly, the subjects with non-VMF frontal damage were also impaired on the standard IGT, consistent with other work.[17] However, their performance did not improve in the shuffled version, indicating that deficits in processes other than reversal learning (perhaps working memory[18] or attention[19]) contribute to their impairment.

Taken together, these two studies argue that a fundamental deficit in reversal learning underlies the aberrant performance on the IGT of patients with VMF damage. This work also suggests that it may be more useful to characterize

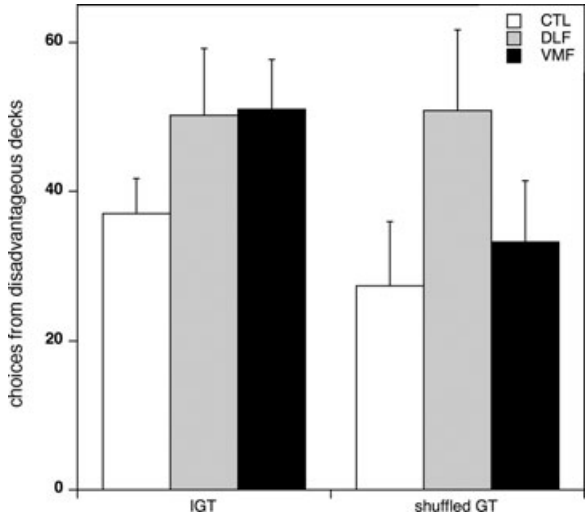

FIGURE 3. Total number of cards chosen from the two disadvantageous decks (over 100 trials) in the original IGT (*left*), and a "shuffled" version of the same task (*right*) designed to eliminate the requirement for reversal learning. Those with VMF damage made poor choices in the original IGT, but performed as well as healthy control subjects in the shuffled version. Those with non-VMF damage (DLF) were equally impaired on the two tasks. Error bars show the upper bound of the 95% confidence intervals. Data from Ref. 16.

the difficulties of such patients as deficits in flexible reinforcement learning, rather than impaired decision making. In a similar vein, others have argued that the particular form of "affective" perseveration captured by reversal learning may also be the basis for the real-life socially inappropriate behavior that can follow VMF damage.[14] This formulation raises other potential avenues of research. For example, while reversal learning is simpler than the IGT, it might be decomposed into still simpler underlying processes. Can the particular process or processes for which VMF is critical be specified in more detail? Finally, can links be made between this learning-based account of VMF function and the neuroeconomics-influenced view that this area is representing value?

Successful reversal learning requires a shift in behavior in response to unexpected negative feedback—either non-reward, or outright punishment, depending on the paradigm. Failure to shift away from punishment in this task also prevents the subject from experiencing (and so learning from) the rewards now available with the alternative choice, even if the ability to learn from reward is intact. One parsimonious explanation of the reversal-learning findings is that VMF is critically involved in adjusting behavior in response to negative outcomes. We found preliminary support for this hypothesis in a study of the effects of VMF damage on the ability to learn from positive and

negative feedback, tested with a probabilistic learning paradigm developed by Frank and colleagues.[20]

Subjects chose between three pairs of arbitrary visual stimuli on the basis of probabilistic positive or negative feedback. Once they met a learning criterion for each pair, they moved to a test phase in which they chose between all possible combinations of the six stimuli, without feedback. Their tendency to choose the stimulus that had been most highly associated with positive reinforcement, and to avoid the stimulus that had been most highly associated with negative reinforcement separately probed their ability to learn from positive and negative feedback. Only five of the 11 VMF subjects (compared with 22 of 24 age-matched control subjects) learned the task to criterion. Consistent with previous work, controls who proceeded to the test phase learned about equally from positive and negative reinforcement. In contrast, those with VMF damage who completed the test phase were markedly and selectively impaired at learning from negative feedback (Wheele and Fellows, manuscript submitted for publication).

Instrumental avoidance learning differs from reward-driven learning in interesting ways: successful learning leads to reduced experience of the negative feedback, which in principle should lead to extinction of the no-longer-reinforced avoidance response. In practice, of course, avoidance learning is not easily extinguished. Various mechanisms have been proposed to explain this phenomenon.[21] Although it is debatable whether probabilistically delivered negative feedback is strictly comparable to outright punishment,[22] the finding that VMF damage disrupted learning from such feedback, while leaving positive feedback-driven learning intact suggests that there are differences in the neural substrates that support the two. It may be that dopaminergic-striatal mechanisms are sufficient for effectively learning from positive feedback, at least in a probabilistic context (and over a few hundred trials), with VMF additionally required for optimally learning from negative feedback over the same time scale.

The links between this reinforcement learning perspective on VMF function, with its emphasis on reversal learning and negative feedback, and neuroeconomic models of VMF as important in representing relative (and typically positively valenced) value are not self-evident. While the work just reviewed argues that VMF is not *necessary* for simple, incremental forms of probabilistic reward-driven learning, it nevertheless may still be involved, perhaps in more subtle or context-sensitive ways (see, e.g., Refs. 23, 24). It may also be that the role played by this region in reinforcement learning paradigms is different from its role (or roles) in other decision contexts. Indeed, even whether the VMF mechanisms important in simple reversal learning are the same as those tapped by probabilistic learning tasks remains to be directly established. Finally, different sub-regions within this relatively large area of the brain may be differentially involved in learning and decision making.

"Relative value" (and the closely related concept of subjective utility) is a powerful construct because it provides a parsimonious mechanism to solve a variety of decision problems. Value is not a fixed feature of a stimulus: it varies according to intrinsic factors, such as satiety, and extrinsic factors, such as the value of other available options. Value can be adjusted for uncertainty or delay and provides a common currency for comparing very different kinds of options and for calculating the total worth of options that have multiple attributes. Determining whether relative value is represented, as such, in the brain is obviously a central problem in understanding the neural basis of decision making. This question has been addressed by functional imaging studies in humans and by electrophysiological work in non-human primates. Both methods have provided evidence that activity in OFC and/or medial PFC reflects relative value, providing an "accounting" that can incorporate many of the factors described above.

This work, reviewed in detail elsewhere in this volume (Padoa-Schioppa, O'Doherty, Wallis), leaves open whether the information about value that seems to be represented in VMF is *necessary* for decision making. This question is most directly answered by loss-of-function methods, such as lesion studies. If we accept that the IGT is detecting the role of VMF in learning, rather than decision making, then the evidence that VMF plays a critical role in decision making is relatively limited. A handful of studies using other gambling or risk paradigms have shown that VMF or orbitofrontal damage can affect decision making in the absence of the need for new learning.[25–27] However, it is not clear whether this effect is specific to risky decision making, or reflects a more fundamental difficulty in determining relative value.

In order to explore these issues further, we examined the effects of VMF damage on a simple form of decision making that involves comparing the relative value of choices in the absence of risk, ambiguity, or trial-by-trial learning. Adapting a paradigm first used in non-human primates,[28] we asked whether VMF damage in humans would disrupt pair-wise preference judgments. Subjects were asked to indicate which of two stimuli they preferred, or "liked better." Categories included colors, foods, and famous people. Within each category, all possible pairs of stimuli were presented. Since subjective preferences are idiosyncratic, there is no right or wrong answer in such a task. Instead, we examined how internally consistent the choices were for each subject. If a given subject preferred food A over food B, and B over C, that subject should prefer A over C. The choice of C over A would be considered inconsistent. We reasoned that if VMF played a critical role in calculating or representing relative value, then damage there should degrade the ability to make these value-based preference judgments, resulting in an increase in inconsistent choices. As predicted, patients with VMF damage ($N = 10$) made significantly more inconsistent choices than either a healthy control group or a group with damage to the frontal lobes sparing VMF.[29]

This result is consistent with a role for VMF in representing relative value, with this region apparently necessary for even this very simple form of decision making. A deficit in this basic process may also explain the changes observed in more complex, multi-attribute decision making after VMF damage.[30] It seems reasonable to suppose that the additional complexity of risky or ambiguous decisions would magnify any deficit in determining relative value in such patients, although this is a claim that remains to be tested directly. Thus, the poor choices made by VMF-damaged patients in many contexts may derive from a degraded ability to compare the value of decision options. This impairment may result in a higher frequency of poor choices, choices that are less consistently risk-averse than those made by healthy subjects, or at least choices that are less consistent than those the patient might have made prior to his or her brain injury.

Could a deficit in comparing relative value also be at the root of reversal-learning impairments that follow VMF damage? After all, reversal-learning tasks require a series of choices between options with changing values. If VMF supports a common component process underlying both reversal learning and preference judgment, then performance on these two tasks should not be dissociable in patients with VMF damage. In fact, in the 10 patients we studied who completed both tasks, overall reversal learning and preference judgment performance were not correlated.[29] At the individual level, three of 10 subjects with VMF damage were clearly normal in their ability to make preference judgments, and two of these three were either moderately or severely impaired at reversal learning. Of the three subjects with the worst preference-task performance, one had only slight difficulty with reversal learning. These findings provide preliminary evidence that reversal learning and judging relative value can be dissociated in some patients with VMF damage, arguing that they are distinct processes with separable neural substrates (although there was no clear relationship between these dissociable behaviors and lesion location in this small sample). These data require further validation, however, not least because the preference task appears to be less sensitive than the reversal learning paradigm.

The series of studies reviewed here illustrates a component process approach to understanding the role played by VMF in human decision making. Candidate component processes were identified based on studies of OFC function in animals and on economic and psychologic models of decision making. This fundamental work has followed separate streams. The first has focused on the role of OFC in reinforcement learning and implicated the region specifically in particular forms of learning, notably reversal learning. The second, less-developed stream is consistent with a role for OFC in representing the relative, subjective value of potential choices. The work described here argues that OFC plays a necessary role in reversal learning in humans, just as it does in animals. Furthermore, this basic process seems to explain the deficits of patients with VMF damage in the more complex IGT. At an even more basic

level, the reversal-learning deficit that follows VMF damage may rest, in turn, on a specific difficulty learning from negative feedback. However, VMF damage also disrupts preference judgments, simple decisions that isolate the comparison of relative value from other aspects of decision making. This supports the hypothesis that this region of the brain is involved in representing or comparing the relative value of options, thereby playing a critical role in human decision making. More generally, these studies underline that work on basic aspects of behavior in animal models can be a powerful starting point for understanding the neural basis of complex human behavior.

ACKNOWLEDGMENTS

I thank Martha Farah and Elizabeth Wheeler for their substantial contributions to the work reviewed here. Funding was provided by NIH R21NS045074, CIHR MOP-77583, and by a CIHR Clinician-Scientist award.

REFERENCES

1. ACKERLY, S. 2000. Prefrontal lobes and social development. 1950. Yale J. Biol. Med. **73:** 211–219.
2. DAMASIO, A.R. 1994. Descartes' Error: Emotion, Reason, and the Human Brain. Avon Books.
3. LOEWENSTEIN, G.F. *et al.* 2001. Risk as feelings. Psychol. Bull. **127:** 267–286.
4. ESLINGER, P.J. & A.R. DAMASIO. 1985. Severe disturbance of higher cognition after bilateral frontal lobe ablation: patient EVR. Neurology **35:** 1731–1741.
5. BECHARA, A., H. DAMASIO, & A.R. DAMASIO. 2000. Emotion, decision making and the orbitofrontal cortex. Cereb. Cortex **10:** 295–307.
6. BECHARA, A. *et al.* 1997. Deciding advantageously before knowing the advantageous strategy. Science **275:** 1293–1295.
7. FELLOWS, L.K. 2007. Advances in understanding ventromedial prefrontal function: the accountant joins the executive. Neurology **68:** 991–995.
8. MONTAGUE, P.R., B. KING-CASAS, & J.D. COHEN. 2006. Imaging valuation models in human choice. Annu. Rev. Neurosci. **29:** 417–448.
9. O'DOHERTY, J.P. 2004. Reward representations and reward-related learning in the human brain: insights from neuroimaging. Curr. Opin. Neurobiol. **14:** 769–776.
10. SUGRUE, L.P., G.S. CORRADO, & W.T. NEWSOME. 2005. Choosing the greater of two goods: neural currencies for valuation and decision making. Nat. Rev. Neurosci. **6:** 363–375.
11. ROBERTS, A.C. 2006. Primate orbitofrontal cortex and adaptive behaviour. Trends Cogn. Sci. **10:** 83–90.
12. BECHARA, A. *et al.* 1994. Insensitivity to future consequences following damage to human prefrontal cortex. Cognition **50:** 7–15.
13. BECHARA, A., D. TRANEL, & H. DAMASIO. 2000. Characterization of the decision-making deficit of patients with ventromedial prefrontal cortex lesions. Brain **123**(Pt 11): 2189–2202.

14. ROLLS, E.T. *et al.* 1994. Emotion-related learning in patients with social and emotional changes associated with frontal lobe damage. J. Neurol. Neurosurg. Psychiatry **57:** 1518–1524.
15. FELLOWS, L.K. & M.J. FARAH. 2003. Ventromedial frontal cortex mediates affective shifting in humans: evidence from a reversal learning paradigm. Brain **126:** 1830–1837.
16. FELLOWS, L.K. & M.J. FARAH. 2005. Different underlying impairments in decision-making following ventromedial and dorsolateral frontal lobe damage in humans. Cereb. Cortex **15:** 58–63.
17. MANES, F. *et al.* 2002. Decision-making processes following damage to the prefrontal cortex. Brain **125:** 624–639.
18. BECHARA, A. *et al.* 1998. Dissociation of working memory from decision making within the human prefrontal cortex. J. Neurosci. **18:** 428–437.
19. HORNAK, J. *et al.* 2004. Reward-related reversal learning after surgical excisions in orbito-frontal or dorsolateral prefrontal cortex in humans. J. Cogn. Neurosci. **16:** 463–478.
20. FRANK, M.J., L.C. SEEBERGER, & C. O'REILLY R. 2004. By carrot or by stick: cognitive reinforcement learning in parkinsonism. Science **306:** 1940–1943.
21. KIM, H., S. SHIMOJO, & J.P. O'DOHERTY. 2006. Is avoiding an aversive outcome rewarding? Neural substrates of avoidance learning in the human brain. PLoS Biol. **4:** e233.
22. SEYMOUR, B. *et al.* 2007. Differential encoding of losses and gains in the human striatum. J. Neurosci. **27:** 4826–4831.
23. FRANK, M.J. & E.D. CLAUS. 2006. Anatomy of a decision: striato-orbitofrontal interactions in reinforcement learning, decision making, and reversal. Psychol. Rev. **113:** 300–326.
24. SCHOENBAUM, G. & M. ROESCH. 2005. Orbitofrontal cortex, associative learning, and expectancies. Neuron **47:** 633–636.
25. HSU, M. *et al.* 2005. Neural systems responding to degrees of uncertainty in human decision-making. Science **310:** 1680–1683.
26. ROGERS, R.D. *et al.* 1999. Dissociable deficits in the decision-making cognition of chronic amphetamine abusers, opiate abusers, patients with focal damage to prefrontal cortex, and tryptophan-depleted normal volunteers: evidence for monoaminergic mechanisms. Neuropsychopharmacology **20:** 322–339.
27. SHIV, B. *et al.* 2005. Investment behavior and the negative side of emotion. Psychol. Sci. **16:** 435–439.
28. BAYLIS, L.L. & D. GAFFAN. 1991. Amygdalectomy and ventromedial prefrontal ablation produce similar deficits in food choice and in simple object discrimination learning for an unseen reward. Exp. Brain Res. **86:** 617–622.
29. FELLOWS, L.K. & M.J. FARAH. 2007. The role of ventromedial prefrontal cortex in decision making: judgment under uncertainty, or judgment per se? Cereb. Cortex **17:** 2669–2674.
30. FELLOWS, L.K. 2006. Deciding how to decide: ventromedial frontal lobe damage affects information acquisition in multi-attribute decision making. Brain **129:** 944–952.