Failure Modes of the Will: From Goals to Habits to Compulsions?

Quentin J.M. Huys, M.B.B.S., Ph.D., and Frederike H. Petzschner, Ph.D.

> Time present and time past
> Are both perhaps present in time future,
> And time future contained in time past. —T.S. Eliot

Good decisions reflect the past and improve the future. Trouble ensues when the rewards of the past conflict with the goals of the future. New goals often fail to override behaviors reinforced in the past, for instance, when patients with obsessive-compulsive disorder (OCD) try to stop the counting behavior that has been their main strategy to avoid aversive intrusions.

Decision-making theories that emphasize the existence of two decision systems working in parallel provide one fruitful multidisciplinary route for understanding the origin of this conflict (1, 2): while habits reflect rewards of the past lengthily accumulated over multiple repetitions, goal-directed decisions capture aims based on the current, more malleable understanding of the world (2). In certain disorders, such as those of impulsivity and compulsivity, an individual's explicit aims appear to repeatedly fall prey to his or her engrained behaviors, suggesting an alteration in the adaptive tradeoff between prospective goals and retrospective habits (3, 4).

In this issue of the *Journal,* Gillan et al. (5) add to an impressive catalog of studies on the *im*-balance between habitual and goal-directed decisions in OCD. Roughly speaking, these studies assess how easily we overcome habits—established through repeated pursuit of an old goal—when a novel goal arises. Gillan et al. have previously shown that patients with OCD have a tendency to establish habits faster than comparison subjects when pursuing rewards (6, 7) and, critically, also when avoiding punishments. Gillan et al. (8) first explained to participants that two visual stimuli, say L and R, would be followed by shocks to their left or right wrist, respectively. A third stimulus, S, signaled safety. Participants could avoid the shock to the wrist threatened by the stimulus by pressing a foot-key with the respective foot (e.g., the left key for stimulus L). A training period in which participants repeatedly avoided the shocks predicted for both wrists was then followed by an instructed devaluation, where the shock electrode was demonstratively disconnected from the left wrist. Participants were explicitly instructed that they would no longer be shocked on the left wrist and that their only goal was to avoid the remaining shocks. Habitual avoidance was measured by the tendency to continue pressing the left foot-key when seeing the now *devalued* stimulus L, despite knowing that this effort was unnecessary because the electrode had been disconnected. OCD patients were more than four times as likely to continue avoiding, indicating that they had established stronger habits than comparison subjects over the training period. There was no difference in the understanding of the devaluation, suggesting that avoidance was not driven by a goal-directed response to an overestimated shock probability. In their present study in this issue, Gillan et al. replicate these findings and extend them with data from functional imaging.

The behavioral replication is noteworthy, since it establishes faster habitization as one feature of OCD, suggesting that compulsions in OCD originate from the rapid transformation of goal-directed avoidance responses to obsessions into habits. Furthermore, habitization and autonomic measures of anxiety went hand-in-hand: patients who formed habits showed similar skin conductance responses for valued and devalued stimuli, suggesting that aspects of anxiety could be related to habit formation. And even though habits are components of everyday life, typically lacking a *compelling* feel, they seemed to have a more urgent nature for OCD patients: the extent of habitization correlated with the participants' self-reported urge to respond and also (more tentatively) with OCD symptom severity. OCD is thus one of the disorders accompanied by a shift from prospective goals toward retrospective habits, akin to addiction (7, 9).

OCD is … one of the disorders accompanied by a shift from prospective goals toward retrospective habits, akin to addiction.

Possible origins of this shift are illuminated by the imaging results. Neurally, the goal-directed and habitual systems depend on different substrates (10), and the authors focused their analysis on these. First, they found that the establishment of habits in the subgroup of patients continuing to respond to the devalued stimulus was accompanied by hyperactivity in the caudate. Patients who did not continue to respond after devaluation showed hypoactivity. Since there was also no overall difference between OCD patients and healthy comparison subjects, it is unclear whether caudate hyperactivity merely reflects the habitization differences or

more general disease processes. Furthermore, animal lesion studies and human imaging studies have associated habit development with the putamen—the caudate is relatively more associated with goal-directed control over behavior (10). Hence it is unclear how hyperactivity of the caudate, particularly in the absence of hyperactivity of the putamen, accounts for increased habitization. In conjunction with the confound between habitization and disease, the authors reasonably interpret the hyperactivity as possibly reflecting futile efforts to impose goal-directed behavior to overcome habits. The alternative, that maladaptive goals are the very source of the persistent responding, is hard to square with the lack of difference in understanding.

Second, Gillan et al. also report a neural difference between OCD patients and healthy comparison subjects: during the acquisition of avoidance behaviors prior to devaluation, OCD patients showed an initial hyperactivation of the orbitofrontal cortex that decreased over time, whereas in healthy comparison subjects, the orbitofrontal cortex activations increased over time. Orbitofrontal cortex neurons are known to represent stimulus value, particularly so in goal-directed scenarios (11). An early hyperactivation followed by a gradual decline might be in keeping with a gradual transition from goal-directed to habitual decisions in OCD patients. However, this finding is complicated by a lack of correlate with habitization itself and does not quite match the hypothesis of futile goal-directed efforts in the devaluation phase.

Overall, the authors' conclusion is that the shift toward habitization in OCD is a result of an impairment in goal-directed control. This is in agreement with the human goal-directed system appearing more malleable by state factors ranging from stress (12) and dopamine (13) to trait variables such as IQ (14). To fully establish the suggestion that a shift toward bias is an endophenotype, it will be necessary to further disentangle and understand the antecedents of the tradeoff between goals and habits.

Computational accounts of the balance between the systems suggest either a competitive setup, in which the relative weight of each system depends on its respective reliability (2), or a more cooperative interaction, in which the goal-directed system improves upon (15) or even trains (16) habits. The intact understanding of the task contingencies and the lack of activation in neural structures associated with uncertainty-based competition (17) speak against the first option. Conversely, the initial orbitofrontal cortex hyperactivation makes the training of habits by internal repetitions within the goal-directed system a distinct possibility. An entirely different explanation for continued responding after devaluation is that sudden changes can be encoded as novel states, which accounts for a broad variety of learning effects (18). A trait-like tendency to avoid generating new states—likely involving the orbitofrontal cortex (19)—would lead to a failure to encode the change and not only explain continued responding after devaluation but also why extinction becomes more efficient with increasing symptom severity (20) despite being overall impaired in OCD.

The further integration of clinical and computational approaches opens the door to a more quantitative understanding of the failure modes of the will in OCD. Clinically, the study by Gillan et al. matches the intuition that, for OCD patients, explanations alone do not suffice. Healing comes from repeated experience and the establishment of novel habits. This, though, is a long and tricky path because the habitual system has to lengthily recreate a novel history such as to align past experience with current goals.

AUTHOR AND ARTICLE INFORMATION

From the Department of Psychiatry, Psychotherapy and Psychosomatics, Hospital of Psychiatry, University of Zürich, Zürich, Switzerland; Translational Neuromodeling Unit, Institute of Biomedical Engineering, University of Zürich and Swiss Federal Institute of Technology (ETH), Zürich, Switzerland.

Address correspondence to Dr. Huys (qhuys@cantab.net).

The authors report no financial relationships with commercial interests.

Accepted December 2014.

Am J Psychiatry 2015; 172:216–218; doi: 10.1176/appi.ajp.2014.14121502

REFERENCES

1. Dickinson A, Balleine B: The role of learning in the operation of motivational systems, in Stevens' Handbook of Experimental Psychology, vol 3. Edited by Gallistel R. New York, Wiley, 2002, pp 497–534
2. Daw ND, Niv Y, Dayan P: Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat Neurosci 2005; 8:1704–1711
3. Everitt BJ, Robbins TW: Neural systems of reinforcement for drug addiction: from actions to habits to compulsion. Nat Neurosci 2005; 8:1481–1489
4. Robbins TW, Gillan CM, Smith DG, et al: Neurocognitive endophenotypes of impulsivity and compulsivity: towards dimensional psychiatry. Trends Cogn Sci 2012; 16:81–91
5. Gillan CM, Apergis-Schoute AM, Morein-Zamir S, et al: Functional neuroimaging of avoidance habits in obsessive-compulsive disorder. Am J Psychiatry 2015; 172:284–293
6. Gillan CM, Papmeyer M, Morein-Zamir S, et al: Disruption in the balance between goal-directed behavior and habit learning in obsessive-compulsive disorder. Am J Psychiatry 2011; 168:718–726
7. Voon V, Derbyshire K, Rück C, et al: Disorders of compulsivity: a common bias towards learning habits. Mol Psychiatry (Epub ahead of print, May 20, 2014)
8. Gillan CM, Morein-Zamir S, Urcelay GP, et al: Enhanced avoidance habits in obsessive-compulsive disorder. Biol Psychiatry 2014; 75:631–638
9. Sebold M, Deserno L, Nebe S, et al: Model-based and model-free decisions in alcohol dependence. Neuropsychobiology 2014; 70:122–131
10. Dolan RJ, Dayan P: Goals and habits in the brain. Neuron 2013; 80:312–325
11. McDannald MA, Lucantonio F, Burke KA, et al: Ventral striatum and orbitofrontal cortex are both required for model-based, but not model-free, reinforcement learning. J Neurosci 2011; 31:2700–2705
12. Otto AR, Raio CM, Chiang A, et al: Working-memory capacity protects model-based learning from stress. Proc Natl Acad Sci USA 2013; 110:20941–20946
13. Wunderlich K, Smittenaar P, Dolan RJ: Dopamine enhances model-based over model-free choice behavior. Neuron 2012; 75:418–424

14. Schad DJ, Jünger E, Sebold M, et al: Processing speed enhances model-based over model-free reinforcement learning in the presence of high working memory functioning. Front Psychol 2014; 5: 1450
15. Keramati M, Dezfouli A, Piray P: Speed/accuracy trade-off between the habitual and the goal-directed processes. PLOS Comput Biol 2011; 7:e1002055
16. Gershman SJ, Markman AB, Otto AR: Retrospective revaluation in sequential decision making: a tale of two systems. J Exp Psychol Gen 2014; 143:182–194
17. Lee SW, Shimojo S, O'Doherty JP: Neural computations underlying arbitration between model-based and model-free learning. Neuron 2014; 81:687–699
18. Gershman SJ, Niv Y: Exploring a latent cause theory of classical conditioning. Learn Behav 2012; 40:255–268
19. Wilson RC, Takahashi YK, Schoenbaum G, et al: Orbitofrontal cortex as a cognitive map of task space. Neuron 2014; 81:267–279
20. Milad MR, Furtak SC, Greenberg JL, et al: Deficits in conditioned fear extinction in obsessive-compulsive disorder and neurobiological changes in the fear circuit. JAMA Psychiatry 2013; 70:608–618, quiz 554