PHILOSOPHICAL TRANSACTIONS B

royalsocietypublishing.org/journal/rstb

Review

Cite this article: Redish AD *et al*. 2021 Computational validity: using computation to translate behaviours across species. *Phil. Trans. R. Soc. B* **377**: 20200525.
https://doi.org/10.1098/rstb.2020.0525

Received: 18 May 2021
Accepted: 28 July 2021

One contribution of 16 to a theme issue 'Systems neuroscience through the lens of evolutionary theory'.

Subject Areas:
behaviour, cognition, neuroscience

Keywords:
computational psychiatry, construct validity, circuit validity, cross-species translation, face validity

Author for correspondence:
A. David Redish
e-mail: redish@umn.edu

†Present address: Springer Nature, New York, NY 10004, USA.

THE ROYAL SOCIETY PUBLISHING

Computational validity: using computation to translate behaviours across species

A. David Redish[1], Adam Kepecs[2,3], Lisa M. Anderson[4], Olivia L. Calvin[1,4], Nicola M. Grissom[5], Ann F. Haynos[4], Sarah R. Heilbronner[1], Alexander B. Herman[4], Suma Jacob[4], Sisi Ma[6], Iris Vilares[5], Sophia Vinogradov[4], Cody J. Walters[7,†], Alik S. Widge[4], Jennifer L. Zick[4] and Anna Zilverstand[4]

[1]Department of Neuroscience, University of Minnesota, Minneapolis, MN 55455, USA
[2]Department of Neuroscience, and [3]Department of Psychiatry, Washington University in St. Louis, St Louis, MO 63110, USA
[4]Department of Psychiatry and Behavioral Sciences, [5]Department of Psychology, [6]Department of Medicine - Institute for Health Informatics, and [7]Graduate Program in Neuroscience, University of Minnesota, Minneapolis, MN 55455, USA

ADR, 0000-0003-3644-9072; LMA, 0000-0001-5535-1498; SRH, 0000-0003-0893-5364; ASW, 0000-0001-8510-341X; JLZ, 0000-0003-3719-8498; AZ, 0000-0002-4889-9700

We propose a new conceptual framework (computational validity) for translation across species and populations based on the computational similarity between the information processing underlying parallel tasks. Translating between species depends not on the superficial similarity of the tasks presented, but rather on the computational similarity of the strategies and mechanisms that underlie those behaviours. Computational validity goes beyond construct validity by directly addressing questions of information processing. Computational validity interacts with circuit validity as computation depends on circuits, but similar computations could be accomplished by different circuits. Because different individuals may use different computations to accomplish a given task, computational validity suggests that behaviour should be understood through the subject's point of view; thus, behaviour should be characterized on an individual level rather than a task level. Tasks can constrain the computational algorithms available to a subject and the observed subtleties of that behaviour can provide information about the computations used by each individual. Computational validity has especially high relevance for the study of psychiatric disorders, given the new views of psychiatry as identifying and mediating information processing dysfunctions that may show high inter-individual variability, as well as for animal models investigating aspects of human psychiatric disorders.

This article is part of the theme issue 'Systems neuroscience through the lens of evolutionary theory'.

1. The challenges of cross-species translation

Humans share extensive similarities with other species in their interactions with the environment and in the behaviours they use to achieve goals within those environments. As such, researchers trying to understand human behaviour can use non-human animal models to investigate mechanisms of behaviour. Importantly, some experiments are differentially feasible in different species. We currently have powerful genetic and circuit access in rodents, allowing microcircuit manipulation of their nervous systems. Mice, rats and monkeys allow for the study of large neural ensembles and direct observation of neural representations, as well as manipulations of neural circuits through lesion, pharmacological, chemogenetic and optogenetic technologies. Human

experiments are generally limited to non-invasive imaging methods, except in certain clinical cases. Further, behaviours may not be comparable across species. In practice, the paths to those behaviours are often different. Humans are typically provided linguistic instructions and very limited experience (training and testing) on a given task, while monkeys are often provided months or years of training, and rats and mice are provided days or weeks of training to accomplish their behaviours. Thus, if our ultimate goal is to understand how brains interact with environments to produce behaviour, we must integrate our knowledge across these different types of experiments, and, thus, we need a process to compare experiments across species.

How do we then translate mechanistic knowledge across species? And, for that matter, how can we combine and integrate knowledge across different levels of analysis in non-human animal species? We propose that an understanding of computational processes can help answer these questions, which requires us to move beyond traditional comparisons of validity to establish a new conceptual framework—that of *computational validity*.

For example, let us assume that our goal is to understand why someone would spend a large portion of their weekly paycheck on cigarettes, even though they found their first experience with cigarettes severely unpleasant, and they continue to state linguistically that they know it is ruining their health and wish to quit. To achieve our goal of understanding what drives their addictive behaviour, researchers must delineate the underlying molecular, neurophysiological and computational processes that are fostering the continued action-selection of smoking. We could examine the neuronal circuits underlying addiction in mice, but mice do not smoke cigarettes, even though they will self-administer nicotine. More importantly, a mouse cannot linguistically tell us whether it wishes to quit smoking, or whether it knows it *should* quit smoking. However, mice can show hesitation or caution, and can re-evaluate behaviours after taking an action, even in the absence of new information. These behaviours suggest that they may experience motivational conflict. This underlying cognitive process of motivational conflict, therefore, may be more fruitfully tested across species, provided it is appropriately operationalized. Importantly, operationalizing the process requires understanding the computations that go into the recognition of and resolution of motivational conflict. Comparing those computations across species brings us to computational validity.

2. Translation and validity

When comparing experimental studies across species, theoreticians talk of at least four kinds of validity [1–6]. (i) *Predictive*, *treatment* or *criterion validity* asks whether an instrument or task reliably predicts a similar measure or outcome across conditions [7–9]; (ii) *face validity* asks whether two behaviours appear intuitively similar [5,10]; (iii) *mechanistic validity* (in neuroscience, often identified as *circuit validity*) asks whether identical neurophysiological or other mechanisms align between experiments or observations [11]; and (iv) *construct validity* seeks to align a theoretical description of an abstraction with the experimental observations seen in different conditions [1,12–16].

We argue here that these concepts of validity are incomplete, and that they miss the important question of computational validity—whether the information processing used during a given behavioural task generalizes across different experiments (species, behaviours, scenarios). Computational validity interacts with these other concepts of validity. We discuss this in more detail below.

(a) Predictive validity

Predictive validity [4–9,16] assesses whether a task or other measure is effective in predicting an outcome. Sometimes referred to as criterion validity [7–9,16], predictive validity includes the concept of treatment validity [7–9,17], which asks to what extent an experimental paradigm (task, measurement) is effective in predicting response to treatment. Although, in a sense, predictive validity is the ultimate goal of preclinical experiments, it has had limited success [4,5,16]. We argue that this comes in part because predictive validity makes no actual claims as to the mechanism of the action, only that it produces similar outward results.

Achieving predictive validity is particularly challenging in neuroscience due to the brain's complexity and the further complexity of the brain's interaction with its environment [18,19]. Additionally, due to the vast social, ethological and environmental differences between species, comparisons of the treatments themselves can be particularly difficult. For instance, successful human treatments for addiction often include methods to substitute another human's judgement during moments of temptation, such as calling one's Alcoholics Anonymous sponsor before drinking [20]. These types of complex social–interactional interventions are difficult if not impossible to model in non-human animals. However, there are examples of predictive success, even within the social realm, such as the social attachment work of Harlow and Bowlby [21–23] or the clear evidence that social isolation increases the susceptibility to addictive drugs in rodents as well as humans [24,25], and that the presence of an alternate option, such as interacting with a conspecific, can reduce self-administration of addictive drugs in rodents [26].

Predictive validity defined on its own has limitations in that it does not attempt to assess the underlying hypothesized mechanisms or constructs of a phenomenon. Given the multifaceted nature of causality, predictive validity depends on more nuanced measures of similarity.

(b) Face validity

Face validity [1,5,10] is generally evaluated based on superficial behavioural similarities and has been successfully applied to basic behaviours that span species, such as a freezing response to an acute threat. Most mammals, including both human and non-human animals, will freeze in response to sudden danger in similar ways [27–29]. As another example, non-human mammals tend to self-administer the same chemicals that humans do, when given the opportunity, even if the methods of delivery differ (smoking a crack pipe versus lever-controlled intra-jugular infusions) [30]. However, taken to its extreme, the invocation of face validity can lead to absurd conclusions, such as the suggestion that humans should poke their nose into ports or that rats cannot be addicted if they cannot indicate their desire to quit, or can only show emotional conflict if they can linguistically describe it to the experimenter.

In addition to these obvious dissimilarities across species, two behaviours that on the surface appear superficially similar could be driven by substantially different mechanisms across species or conditions. As an example, the activity-based anorexia rodent model of anorexia nervosa is predicated on the assumption that the specific mechanisms that underlie the excess weight loss of rodents' wheel-running to the exclusion of food consumption are the same as the mechanisms that underlie the excess weight loss seen in individuals with anorexia nervosa [31]. Humans, however, experience a host of different social and environmental drivers for weight loss goals [32], while there is no evidence for the same social environment precipitants for weight loss in rats.

To get around this difficulty of superficial relationships, one must consider other types of validity, including mechanistic, construct, and, we argue here, computational validities, that address the underlying components of the behaviour.

(c) Mechanistic validity

Mechanistic validity [11], often referred to in neuroscience as circuit validity, measures the degree to which underlying mechanisms or homologous neural circuits across species are involved in similar ways during a given behaviour. For example, the identification of dopaminergic dysfunction as a critical mechanistic step in Parkinsonian behavioural dysfunctions supports experimental dopamine depletion in monkeys or rats as a means to understand the underlying neurophysiological dysfunction contributing to Parkinson's disease [33–35]. Mechanistic validity asks whether dopaminergic manipulations have similar effects on downstream structures across species, even though the mechanism by which dopaminergic dysfunction is induced is different in animal models as compared to humans with Parkinson's disease. Experimental comparisons based on mechanistic (circuit) validity will be more likely to produce good predictive (treatment) validity.

Circuit validity works particularly well for conditions in which there is a strong homology between species. For example, the amygdala plays a central role in defensive responses across species, such as freezing in rodents and increased galvanic skin responses in humans to cues predicting punishment [36]. However, circuit validity becomes difficult when there are disagreements about homologies and when there are clear differences in circuit anatomy between species. For example, the homologies between rodent and primate prefrontal cortices are deeply controversial, making circuit validity difficult if not impossible for some questions [37–41]. For instance, whether rodents have a homologue of the primate dorsolateral prefrontal cortex, a key structure for executive control in humans, is very much under dispute. Primates (including humans) may be more cortically dependent than rodents in general [42], which likely changes the underlying functionality of many circuits. To circumvent these issues, experimentalists often focus on abstractions of the overall construct they are attempting to study, in order to provide translation across species.

(d) Construct validity

Construct validity [1,12–16] assesses the degree to which an experimental design will provide observations that align with a theoretical model of a specific behavioural, psychological or cognitive process.[1] It builds on the understanding that different behaviours may reflect the same underlying construct. For example, one experiment might ask a human to remember a number told to them linguistically and then repeat it back after a few minutes [44–46], while another experiment might show a monkey a pair of objects, hide them, and then reveal them again [47,48], asking them to move the object to reveal a reward, and another experiment might ask a rat to return to a location previously experienced for reward [49,50]. These experiments access three entirely different behaviours across the three species, but all require the subject to remember a piece of concrete information across a time gap and thus theoretically access the construct of working memory [51–53].

(e) Computational (or algorithmic) validity

Our contention is that all of these validity considerations are important and that animal models should address all of them. However, we argue that there is a critical, but often missing, validity comparison that exists alongside these other validities and potentially integrates them in an important way: that of computational validity.

We define 'computation' here as a formal process addressing how information is stored and processed within an agent performing a task. We include within our term 'computation' both a description of the task-relevant information that must be represented in order to achieve a task goal and also a description of the algorithmic processes by which this information is encoded and manipulated. These form a continuum of formal description that can be used to compare questions across tasks and species. One of the most important discoveries in the computational sciences over the last 50 years is the observation that how one represents data shapes how one can efficiently process it, and furthermore that how one processes that data can change the behavioural consequences of a task [18]. As such, the question of computation is one of what information is represented within the system, how that information is transformed and made available to other processes within the system, and how that information is used to guide behaviour.

A given task can be described at multiple levels of abstraction that provide different predictions with different granularities and specificities. While distinctions are often made between 'computational' and 'algorithmic' descriptions [54], we include measures of similarity and dissimilarity of both of these in the term 'computational validity'.

In addition to the behavioural neuroscience examples used in this manuscript, computational analyses can be applied to multiple levels of abstraction within a neural system (subcellular, single cellular, network, cognitive). Questions of information processing can be applied at all of these levels. Our focus here is on behaviour and our examples are high-level, but lower-level computational analyses also have behavioural consequences. For example, retinal receptors responding to specific wavelengths of light, colour being measured through an opponency process, and visual cortex cells normalizing firing to ambient levels, are all low-level computational processes that have been important to our understanding of visual perception and critical in our ability to translate discoveries across species. The question of levels of abstraction is beyond the scope of this paper, but has recently been discussed in detail elsewhere [55] and warrants further consideration in the investigation of computational validity.

Fundamentally, if behaviour depends on information processing in the nervous system, then the key to translating

between observed behaviours across different species is to align the computational and algorithmic processes that underlie the behaviours, asking (i) what information is being represented and maintained (versus discarded/ignored), (ii) how is that information being processed (algorithm) and (iii) how do specific behaviours (output) arise from that information processing cascade. We argue that the key to computational validity is to first operationalize behavioural or cognitive processes as a computational process. That is, rather than trying to define a cognitive process such as 'working memory' without defining its underlying computations, we need to identify the computational steps that we commonly describe as working memory operations [50,51,56,57]: what are the inputs that enable or trigger working memory computations? What are the potential outputs (stored information—complete or partial or contaminated by distractors)? What are the steps in the computations? And what are common failure modes that correspond to the different steps in the computations? These questions differentiate storing information across a time gap and processing that information. Moreover, if we take neural populations as performing a computation, then that computation can resolve differently in different behavioural tasks [45,50,58–60], and various hypothesized computational processes will predict contrasting patterns of behaviour and neural activity across those tasks. Dissociating the relevant computations becomes critically important for resolving the underlying neural circuit mechanisms.

For example, one can define deliberation and planning as entailing an explicit imagination of a potential outcome, an evaluation of that outcome, potentially in the light of other remembered options, and then a decision based on those deliberations [18,61–64]. This process is fundamentally different from that of procedural, cached action chains, in which one recognizes a situation and releases a well-practiced action chain [18,65,66]. Neural studies of the hippocampus in both rodents and humans have provided evidence for the construction of those imagined outcomes in deliberation/planning, most likely instigated by inputs from the medial prefrontal cortex [67–71], while dorsolateral striatal neural circuits learn to represent situation–action pairs useful for procedural decisions, but do not contain information about those future outcomes [50,72–74].

3. The complexity of behaviour (getting the ethology right)

Asking what underlying algorithm is being used is particularly important because animals (including humans) are not general information processing machines, but rather carry out their behaviours within their species-specific ethological limitations. This means that it is critical to 'get the ethology right'. For example, it is often easier to ask primates (humans, monkeys) to categorize visual signals, but easier to ask rodents to categorize olfactory, auditory or spatial signals. Thus perceptual decision-making has been studied through the categorization of random dot motion (are most of the dots moving left or right?) in primates [75–77], but by using clicks (frequency or side) or through running past spatial cues in a virtual environment (number of 'posts' on the left or the right) in rodents [78–81]. While these signals can arrive through different sensory modalities in the different species, homologous cognitive structures and information processing operations are evoked [82–87].

In general, species have evolved ethological processes that make some behavioural domains easier to access than others. Thus, when asking questions about computationally similar processes, we may need to reveal those processes through ethologically designed tasks. For example, on lever-press experimental paradigms, rats tend to perseverate (defaulting to win-stay algorithms), while on spatial experimental paradigms, rats tend to alternate (defaulting to win-shift algorithms) [50,88–90]. If we want to study economic decision-making processes in rats, we need to provide them with environments in which they will reveal those processes [91–94]. Importantly, these observations are not limited to non-human animals. Humans also find it easier to identify the counter-positive in logical puzzles if framed in a cultural manner that they have experience with [95].

(a) The problem of equifinality (similar behaviours can arise from multiple algorithmic processes)

One of the key challenges for behavioural experiments is that a given behaviour can arise from multiple algorithmic processes. To see how multiple decision-making algorithms can produce a given action, one can look at the classic plus-maze task. In this task, rats are exposed to a plus-shaped maze and then trained to run from the south arm to the west arm. There are two computational processes that rats could be using to solve this task—they could use a representation of the spatial relationship between start and goal to plan a path, using a cognitive map, or they could learn to associate being put on the maze with turning left [50,96–99]. As will be laid out in depth below, these two representations engender very different computational processes: knowing the spatial relationship between start and goal enables a search process that hypothesizes the consequence of one's actions, allowing the evaluation of that consequence, which enables flexible but slow action decisions [18,60,66,100–102]. By contrast, an association between the maze and the action requires only a recognition of the situation and the release of the associated action, enabling fast but inflexible action decisions [18,60,66,97,102,103]. Turning from the south arm to the west arm on the plus-maze cannot differentiate these computational processes, but a probe trial in which the rat is placed on the north arm can. The cognitive map strategy from the north arm will reach the west arm by turning right, but the situation–action association will turn left, taking the rat to the east arm. On this task, rats normally transition from cognitive map to situation–association strategies with experience [99,104].

A similar decision-making transition has been seen in other tasks, particularly sequential tasks wherein a subject can get into a flow, but that flow can be disrupted. For example, in the left–right-alternation task, a rat learns that there are three potential contingencies to achieve reward (make a left lap, make a right lap or alternate sides). When contingencies change, cognitive map processes drive behaviour, but when an animal runs a single contingency for a number of laps, procedural systems begin to drive behaviour [72,105–107]. Primates show similar effects in the telephone task, in which subjects have to enter a sequence on a grid of numbers [108–111], and humans show similar effects in the serial reaction-time task, in which the subject is given a keypad

of buttons and told to push the button that lights up. Unbeknownst to the subject, the buttons light up in a sequence. With extensive experience with the sequence, behavioural control transitions to a smoother procedural process [108–115].

The equifinality seen in these tasks demonstrates how computational validity can be used to ascertain circuit validity [18,63,65,66,72,104,110,112]: early actions on these tasks are driven by explicit, deliberative decision processes that neurophysiologically include representations of goals, paths and outcomes, but create variable paths in action execution. By contrast, after repeated exposure, late actions on these tasks are driven by implicit, procedural action-chain processes that neurophysiologically include representations of the situation–action relationships. Early actions are more flexible, but also slower and more variable in their execution. The different decision-making models (a forward-looking planning process versus a backward-chained reinforcement learning process) produce subtle variations in these tasks that change as animals transition between these two decision-making processes. Moreover, early learning produces flexible behaviour, while late actions are less flexible, but also faster and more reliable in their execution. Early processes depend on interactions between the hippocampus, prefrontal cortex and medial striatum, while late processes depend on the motor cortex, cerebellum and dorsolateral striatum. Even though the plus-maze, the left–right-alternate task, the telephone task and the serial reaction-time task are superficially different, they access similar computational processes, involve parallel circuits and show similar predictive validity (figure 1).

The long history of literature exploring these issues has found that because some behavioural measures show similarity (e.g. error rates can be low whichever system is driving behaviour), it is critical to measure multiple aspects of behaviour simultaneously and apply manipulations to constrain the algorithms. For example, probe trials in which the rat starts from the north arm produces different outcomes under deliberative/planning strategies and procedural/habit strategies [98,99,104]. Furthermore, the quantitative assessment of behavioural execution can reveal subtle differences characteristic of distinct algorithms. Deliberative processes show more variability in the paths taken, including pause and re-orientation behaviours at choices, and hesitation at components [63,109,120,121]. Procedural processes permit anticipatory motor preparation, which leads to smoother navigation paths in rats and smoother finger paths in monkeys [72,106,108,111,121].

Importantly, because these computational hypotheses depend specifically on information processing, neurophysiological measurements that directly measure the information within neural systems (such as the decoding of neural signals and changing tuning curves) can test those predictions [122]. For example, in the tasks shown in figure 1, one can directly observe different information processes in the hippocampus, ventral striatum and dorsolateral striatum [63,72]. Hippocampal ensembles sweep representations of location from the current position of the rat to the goal [67,123], also seen in similar tasks [68,69,124–126], consistent with a hypothesized role in planning. Ventral striatal signals show transient representations of goal outcomes during early learning [127] and ramps of increasing firing during late learning [128], consistent with a hypothesized role in evaluation. By contrast, dorsolateral striatal ensembles develop bursts of firing at the start and end of the journey as the behaviour automates [105,129–131], and cells that encode different actions to be taken at different points of the maze [105,129,132–134]. Similarly, one can find quantitative signals in parietal cortices that integrate information as predicted by drift–diffusion and race-to-threshold models [87,135], and confidence-related signals in the lateral orbitofrontal cortex [136,137].

4. Examples of uses of computational validity

As noted above, while humans and other animals use computational abilities to guide adaptive behaviour, they did not evolve as general computational machines. This means that when asking questions about abstract computational abilities, it is important to design tasks that access the inputs and abilities of a given species. It is also important to provide each species with an appropriate output that can be used to reveal the computational process. For example, if one defines addiction as continued costly behaviours despite a stated preference to stop the behaviour, then non-linguistic animals (such as mice or rats) can never be addicted. However, it is possible to identify motivational changes, for example, an increased willingness to pay a cost for drug delivery, and to measure motivational conflict within mice and rats through hesitation and re-orientation behaviours, both of which are increased in subsets of animals willing to pay high costs for drug delivery.

(a) Perceptual evidence accumulation

An early example of the utility of focusing on computational validity has been accumulation-of-evidence models [82,87,138,139]. This class of models proposed an explanation for how observed patterns of choice accuracy, reaction times and evidence are related through a simple computational process: evidence is accumulated over time until a threshold is reached, releasing a response. With the earliest models, it was possible to separate when the individual's accumulation process began, how rapidly they accumulated evidence, what their evidence threshold was and whether they had a bias towards a type of response [140], each of which have provided measurable targets in neurophysiological recordings in non-human animals. Precisely defining the exact noise process, how evidence is integrated, and other necessary components resulted in the ability to parse unique differences that could match a range of behavioural patterns in humans [141,142].

These models imply that evidence accumulation should depend on the information available at each moment—for example, in the random dots task, accumulation should depend on the coherence of the stimulus [143]. In this task, participants are shown a large number of dots on the screen that move in a random direction except for a proportion of these that move in a particular direction. The dots that are cohesively moving in a particular direction indicate which response the participant should give. By manipulating the proportion of dots moving in the same direction (i.e. the overall coherence of the representation) the experimenter can manipulate the rate of information provided. Typically, individuals show slow reaction times in trials that have low stimulus coherence and faster reaction times to trials with high stimulus coherence, which can be quantitatively predicted based on evidence accumulation [76]. There are a number of related algorithms for evidence

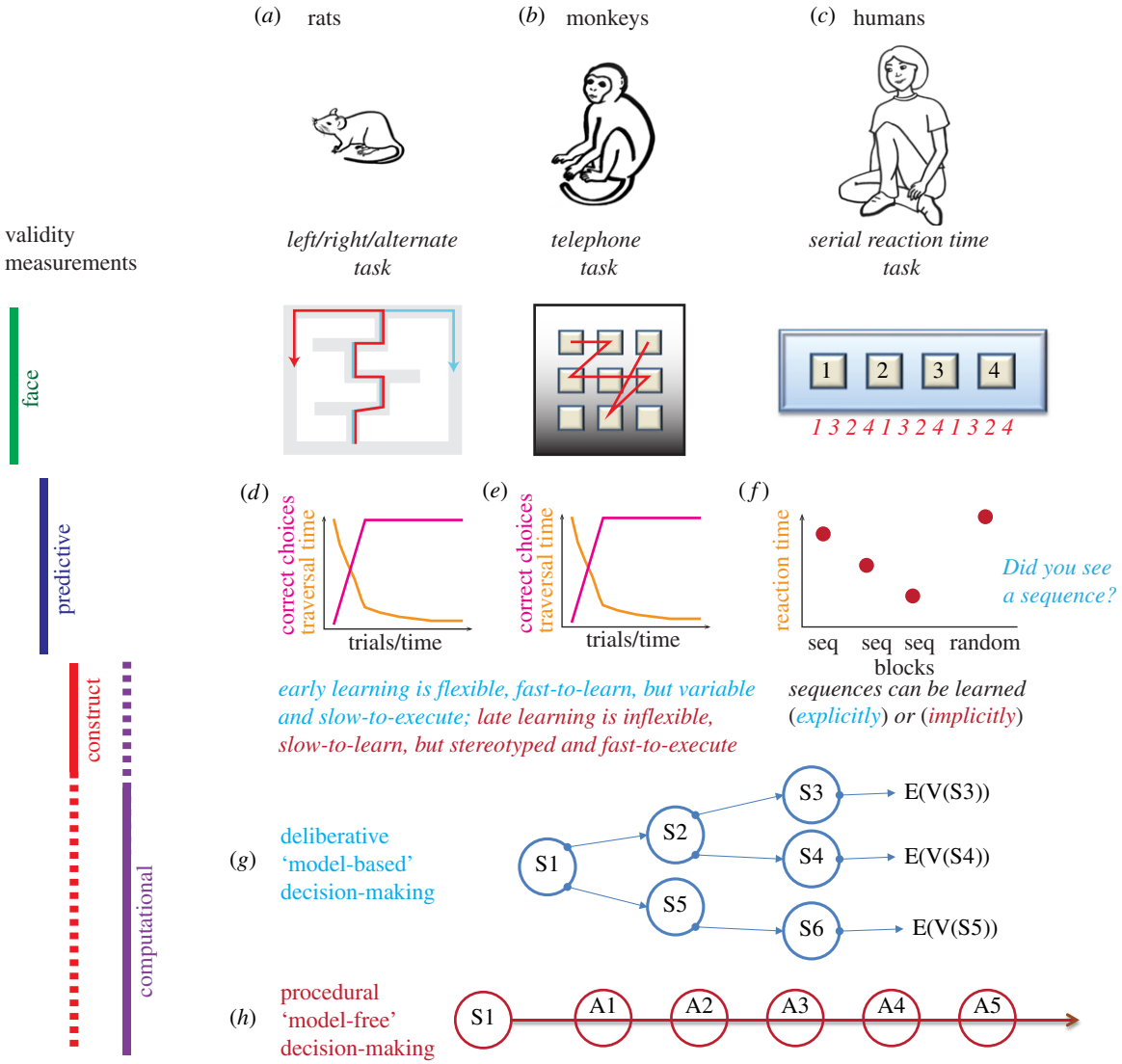

Figure 1. An example of different kinds of validity measures applied to cross-species tasks. (*a*) Rats run the left–right-alternate (LRA) task, in which they learn to run left for food, right for food, or alternate, and then change contingencies one or more times in a session [106,107,116,117]. (*b*) Monkeys learn to type a sequence of keys on a telephone keypad [108–111]. (*c*) Humans are told to push a button that lights up; unbeknownst to them, the buttons light up in a sequence [113,115]. (*d*,*e*) Both the LRA and telephone task show an early fast learning followed by a slower automation of behaviour that does not increase the correct choices, but is characterized by an increase in anticipatory movements, particularly in monkeys. Early learning is dependent on and reflected in the information processing of prefrontal, hippocampal and ventral striatal components, while the later automation is dependent on and reflected in the information processing of dorsolateral striatal components [64,72,104,106,110,118]. (*f*) Similarly, humans show both decreases in reaction time (and increases in anticipatory movements) and a dissociable declarative recognition of the sequence. The motor changes are reduced in patients with Parkinson's or Huntington's disease, while the declarative sequence recognition is disrupted in patients with Alzheimer's disease [113–115]. Theories explain the early learning as deliberative, declarative and explicit and the late learning as procedural and implicit [46,50,60,72,102,119]. Computationally, theories describe early learning as dependent on 'model-based' search processes that include explicit representations of outcomes (*g*) and late learning as dependent on 'model-free' action chains (*h*) [18,65,72,112]. *Face validity* measures direct task similarity. *Predictive validity* measures whether similar manipulations appear across the tasks, such as, for example, the effects of different neural disruptions. *Construct validity* measures the theoretical components of the task. *Computational validity* measures the extent to which the tasks are solved by similar computations. (Online version in colour.)

accumulation, from drift–diffusion to race models [144], and identifying which specific variant can explain specific behavioural patterns remains a challenge [136,145,146].

Neurophysiological recordings in non-human primates have identified neural signals that encode the evidence as it accumulates, and which accumulates from a starting point reflecting the subject's experience (the bias), accumulating faster for more coherent signals, and reaching a consistent threshold before the choice is initiated [75,87,146,147]. Rodents have poorer visual ability compared to the primates, so researchers have either significantly reduced the visual complexity of perceptual evidence accumulation tasks or have changed the sensory modality to use auditory, olfactory or tactile cues [78,79,136,148,149]. But after making these changes to sensory modalities, neural and behavioural correlates of the drift–diffusion process were found in rodents [150]. Careful analysis of the behavioural data within a computational framework has enabled these models of perceptual evidence accumulation to relate task performance neurophysiology across species.

(b) Fear conditioning and anxiety

Classical fear conditioning experiments are actually built primarily on face and circuit validity rather than construct or computational validity. For instance, these experiments

largely capitalized on cross-species similarities in biobehavioural responses to acute (and perceived) threats that were tractable in laboratory settings (face validity) [29,151–154]. Neurophysiologically, the likelihood of observing the response is both causally and correlationally related to the synaptic efficacy across the lateral and basolateral to central amygdala connection (circuit validity) [36]. Behavioural experiments that create conditioned responses also lead to increased synaptic efficacy across the amygdala [155]. Manipulations of the strength of that connection change the performance of that response [156].

While these models were often justified as a means to address questions of post-traumatic stress disorder or anxiety, they have not been particularly successful as such [152,157,158]. The development of cross-species tasks that capture the human experience of anxiety has proven challenging [5,159,160]. Early theories of anxiety (such as [161]) included underlying computational hypotheses—that anxiety entailed imagined representations of future threat and underlying conflicts between approach and avoidance motivational goals [162], but these computations are not easily accessed through classical fear conditioning experiments [152,157,158]. Computational analyses of threat assessment that take into account distance between predator and prey and the available action strategies may provide a more translatable story [154,163,164], but have not been as well explored computationally. However, new tasks that directly access approach–avoid conflict, hesitation and representations of potentially dangerous future outcomes may provide more direct access to these computations [63,164–167]. Direct manipulation of outcome uncertainty may provide additional opportunities for identifying specific computations [168–170].

Experiments in which rodents hesitate before taking actions that may lead to threat have been suggested as a more computationally valid measure of worry and anxiety [154,164,165], including classic observed behaviours such as the stretch-attend posture, seen before progressing out from safe to dangerous zones [63,171–173]. Neurophysiological studies suggest that these moments may include imagined representations of future threat [174–176], suggesting computational validity, while pharmacological manipulations producing similar outcomes may suggest predictive validity as well [166,173]. We argue that applying computational validity measures to the underlying processes within these tasks is likely to improve translatability beyond the simpler paradigms that have dominated the classical literature.

(c) Restaurant Row/WebSurf

In the Restaurant Row task, mice or rats forage for differently flavoured food in sequential encounters with four 'restaurants' [91,92]. In the WebSurf task, humans forage for videos in sequential encounters with four 'galleries' [177–179]. Although the modalities are different, both rewards are consumed in task, and it has been possible to align the computational decision flow between the two tasks, revealing computational similarities between the species [93]. In general, on encountering a restaurant/gallery, a delay is revealed, and the subject is given the option to wait out the delay for reward or to skip the reward and proceed on to the next restaurant/gallery. Subjects are time-limited and thus are spending time from a budget to maximize their reward intake. Mice, rats and humans all typically exhibit thresholds for each restaurant/gallery, such that if the delay is lower than that threshold, they wait out the delay, but if higher, then they forgo the reward. For mice and rats, receiving many days of experience, these thresholds are stable—differing from animal to animal and from flavour to flavour, but remaining constant over days [91,92]. For humans, these thresholds are consistent with stated preferences, including rankings of the four galleries, and average ratings of individual videos [177–179]. These observations are consistent with the concept that each restaurant or gallery provides a reward of a given subjective value and subjects wait out the delay if the cost is lower than that value.

In versions of the task with separate offer and wait zones, where delay is revealed, but does not count down in the offer zone, and only counts down on entry into a wait zone, the normative behaviour would be to proceed through the offer zone to the wait zone, make the decision in the wait zone and quit if necessary. Neither mice, rats, nor humans behave in this manner. All take time to make decisions in the offer zone, and show a resistance to quitting out of the wait zone [93]. Other computational similarities can be seen between these two tasks. For example, reaction times in the offer zone are increased not at the threshold itself (as would be expected from a simple perceptual model), but rather just above threshold, suggesting that subjects find it easier to stay than to reject a given offer where the cost and value are close [91–93]. From this, we conclude that human and non-human animals are likely using similar decision-making processes in these two tasks, even though the two tasks access different perceptual–action modalities. Interestingly, a subset of humans do show reaction times peaked at thresholds, suggesting that this subset of humans may be using a different decision-making process [180].

(d) Measuring decision confidence across species

Historically, confidence judgements have been taken to be a prime example of a uniquely human cognitive capacity, metacognition, which would make confidence unsuitable for translational studies [181–183]. While intuitively the sense of confidence reflects a process of apparent self-reflection, it can be tremendously useful for survival in an uncertain world. Determining how much time or effort to invest, whether in the stock market or a rich food patch, requires accurate estimates of confidence about each option. Indeed, confidence can be also defined as a statistical quantity, the likelihood that a belief is correct [184,185]. This definition lends itself to a computational operationalization with the potential to connect behavioural observations across species. The key insight is that we can create behavioural tasks that incentivize the use of decision confidence so that making confidence-guided choices somehow benefits the subject. Incentivized subjects can demonstrate that these confidence-related computations drive behavioural strategies.

Using this approach, confidence-guided behaviours have been shown in rats, mice and non-human primates [136,186–190]. For instance, rats were trained to first decide between two options based on noisy sensory information and then to wait for uncertain delayed rewards. The rats' time investments from trial to trial quantitatively matched the statistically appropriate use of confidence information, which can be inferred based on their choice behaviour [191,192]. In this case, the use of a confidence computation can be determined based on a normative statistical theory. Similarly, experiments

that have asked monkeys to wager rewards or opt out of choices [186–188,193], or have allowed preverbal infants to ask for help or persist in choices [194] have found that they behave in proportion to statistical confidence. There are also numerous algorithmic models that have been used to explain confidence-guided behavioural strategies and learning across species, including reinforcement learning, statistical classifiers and evidence accumulation models [136,186,195–198]. Importantly, explicit self-reports of subjective confidence in people have also been found to reflect these statistical computations [199,200]. Based on these approaches, there is increasing neurobiological understanding about the brain regions supporting confidence, including single neuron and inactivation studies in orbitofrontal, frontopolar, anterior cingulate and parietal cortices, as well as the pulvinar and supplementary eye field regions [137,189,201,202].

Dysfunctions of confidence appear to contribute to a range of psychiatric disorders and have been found in numerous clinical patient populations [203,204]. For instance, underconfidence is associated with pathological doubt in anxiety disorders including obsessive–compulsive disorder [205–207]. Overconfidence is a characteristic of narcissistic personality disorder [208], while patients with major depression tend to exhibit attenuated and biased confidence reports [209,210]. These studies are increasingly yielding quantitative metrics in behavioural tasks that can be assessed in both human patients and in non-human animal models.

5. Models of computational change

Computational validity is a particularly useful construct when we examine treatments and their effects.

An interesting example lies in the common addiction treatment of Contingency Management, in which subjects are rewarded (monetarily) for not succumbing to their addiction in a recent time frame (for example, not using their drug of abuse for the previous week) [211,212]. Early theories of Contingency Management were based on hypotheses of alternate reward and increased economic opportunity costs, but Contingency Management works better than would be expected given this theory [213,214]. If one measures the expected effect of the rewards offered in Contingency Management given the elasticity of drug use as a function of cost on the street, Contingency Management works much better than expected [215,216].

Rats make different valuation decisions when faced with breakpoint experiments (how much effort is an animal willing to expend to receive a drug?) than when faced with choice experiments (which of two options would an animal prefer?) [217]. Computational theories suggest that these two experiments access different decision-making system processes, consistent with human economic studies suggesting a difference between willing-to-pay experiments and choose-between experiments [18,218]. Regier & Redish [216] suggested that Contingency Management may be computationally akin to these changes in experimental paradigms, providing the addict with a deliberative choose-between option which interferes with the willing-to-pay decisions usually made.

Computational theories suggest that decisions about particular futures arise from deliberative decision processes which entail imagination and evaluation of those future outcomes [18,61–63]. This process is referred to as *episodic future thinking*. Rats, monkeys and humans have all been found to neurophysiologically imagine future outcomes using similar processes, including explicit representations of those future outcomes, and through similar neurophysiological circuits (involving ventromedial prefrontal cortex in humans, the homologous medial prefrontal cortex in rats, as well as hippocampus, nucleus accumbens and orbitofrontal cortex in all three species) [63,67,71,219,220]. The explicit nature of the representation necessary for future evaluation suggests that concrete futures are easier to imagine and evaluate than abstract futures. This may be one reason that Contingency Management works so well—it provides a concrete option to look forward to. These hypotheses suggest that Contingency Management will work best with concrete options (rather than simple monetary rewards), that it will depend on prefrontal–hippocampal–accumbens circuits and on an intact orbitofrontal cortex, and that it could be improved with episodic future thinking training [221,222] and motivational interviewing [223,224].

6. When computations do not translate

All of these validities (face validity, predictive [treatment] validity, mechanistic [circuit] validity, and construct and computational validity) are, in actuality, measurements; that is, they ask the question: *to what extent are these two experiments similar or different?* We can learn useful information both from when we find close validity—a strong similarity between the experiments—and from when we find disruptions in validity, when an expected similarity is found instead to be dissimilar. These concepts of validity are particularly important when translation fails. Moreover, these measures interact in important ways. Recognizing differences in circuit validity can be important when treatment validity fails. Recognizing differences in computation can explain differences in how different species (or different subjects) process different constructs.

Early models of navigation assumed that cognitive maps were built by stringing routes (chains of cues) together, but studies found, instead, that representations of allocentric spatial information entailed an internal representation of a coordinate system which cues were then associated with [50,60]. These two different theories make very different predictions when faced with mismatches between external cues and internal coordinate frames. For example, early descriptions of place cells (hippocampal cells which encode location within an environment [225]), and head direction cells (cells in the postsubiculum and anterior thalamus that encode orientation within an environment [226,227]), assumed that these cells were derived from cues (see [50] for a historical review). However, attractor network computational models of the head direction system suggested that the internal coordinate structure of a one-dimensional circular ring (orientation) came first and external cues could be associated with the representation on the ring to reset the representation if the animal became disoriented [228–230]. These models suggested that if the internal coordinate frame changed on each experience, cues would never become associated with a given spatial signal because they would appear as unstable to the rat, because the models suggested that the rat prioritized the internal representation over the external cues. Testing this theory directly, Knierim *et al.* [231] found that the place and head direction cells in a disoriented rat never became tied to

external cues. Gallistel and colleagues [50,232] tested this in a simple behavioural experiment in which rats were placed in a rectangular environment and highly salient cues were provided at the corners. Non-disoriented rats (whose head direction representations were thus consistent on each entry) were capable of learning the cues and identifying one unique corner to gain food reward [233]; however, rats who were disoriented on each entry (thus with head direction representations different on each entry) were unable to differentiate the opposite corners (which were geometrically equivalent) [234].

When Hermer & Spelke [235,236] first tested this in humans by asking people to find an object placed in a coloured box, they found that humans did not show this disorientation effect. But, of course, the humans were able to remember the location of the object across the disorientation by linguistically repeating the description. 'It's in the blue box. Blue box. Blue box…' If they gave the humans a linguistic blocking task (counting backwards by sevens from 1000, for example), then even adult humans reverted to showing the similar disorientation effects that rats did. Similarly, children who did not yet use linguistic orientation words ('to the left of') were unable to differentiate a box close to a black wall from a box far from it after disorientation, but children who had those linguistic orientation words could.

Because the navigation literature had applied computational analyses to the various processes underlying maps, orientation, cues, and how cues reset those cognitive map representations, it became possible to identify how language changed human memory signals, and were able to reveal that hidden under those linguistic mechanisms were similar computational processes in both human and non-human animals.

7. Individual differences

Designing tasks for computational validity might also address a major open problem in clinical neuroscience—reliable measurement of between- and within-individual differences. A common goal of behavioural neuroscience is to understand sources of behavioural variability, and particularly sources of extreme/outlier behaviour that manifest as mental disorders. These disorders are internally heterogeneous—patients with a common diagnostic label such as 'addiction' or 'depression' can report very different symptom patterns [237,238]. There is also remarkable comorbidity between disorders, to the point that multiple diagnoses are more common than 'pure' syndromes [239]. A prominent view, exemplified by the US National Institute of Mental Health's Research Domain Criteria (RDoC) project, argues that the solution to heterogeneity and comorbidity lies in a new, quantitative taxonomy of mental illness [240]. In this framework, it is commonly assumed that patients can be reliably phenotyped by comparing their performance on psychophysical tasks to an appropriate set of norms [237,241,242]. Unfortunately, emerging evidence suggests that standard psychophysical performance metrics (response times, correct responses, number of trials to criterion) are poor measures of inter-individual variability. In fact, they were designed to be poor measures of inter-individual variability, because they were designed to increase the contrast between groups [243–245]. Because we normally want to analyse a task in terms of the contrast between two or more conditions/trial types, tasks and stimuli are generally designed and psychometrically validated to consistently produce differences between conditions.

A computational perspective might recover viable individual-difference metrics even from tasks that are designed to suppress those differences. As noted above, a participant might arrive at a correct response or a series of economic choices by many different algorithms. In the presence of significant circuit dysfunction, some of those algorithms may be inaccessible or disfavoured, and computationally informed analyses might be able to detect these algorithmic biases. For instance, recent studies identified deficits in construction/use of reward contingency models in patients with compulsive disorders, even when those patients showed no outward deficits on a reward learning task [242,246]. While most common decision models have not yet been validated for test–retest reliability, if that work is done, computational analyses might also track changes in response to treatment or might provide biomarkers of successful clinical target engagement [237,247]. For instance, a validated measurement of episodic future thinking might help identify cases where a contingency management intervention was failing to boost that thought pattern.

8. Conclusion

In sum, it is important to always be asking what computations the individual subject may be performing to accomplish a given task. These computations will have consequences that constrain the animal's behavioural responses (perhaps with subtle changes under probing conditions), will make predictions about what information is encoded within different neural circuits (an important step towards measuring circuit validity), can be linked to specific manipulable components (important for finding treatment validity), and can reveal critical inter-individual differences relevant for the understanding of psychopathology.

Monkeys are not small humans. Rats and mice are certainly not. No rat has built a spaceship to the moon (although monkeys were in space before humans and there have been rats navigating mazes in space). It would be ludicrous to argue that all of these species are performing the same computations in all behaviourally similar situations. Rather, we argue that by delineating the computations being performed in a given task, we can identify the similarities and differences underlying the information processing and the behaviour both across species and between individuals of the same species. This should improve our ability to translate knowledge and understanding across species.

Data accessibility. This article has no additional data.

Authors' contributions. All authors co-wrote the paper together.

Competing interests. We declare we have no competing interests.

Funding. P50 MH119569 (general). Minnesota Medical Discovery Team - Addictions (general). T32 DA037183 (O.L.C.). T32 DA007234 (C.J.W.). R21 MH120785 (A.S.W., A.D.R.). K23 MH112867 (A.F.H.). K23 MH123910 (L.M.A.). R01 MH097061 and R01 DA038209 (A.K.).

Acknowledgements. We thank the NeuroPlasticity Research in Support of Mental Health (NeuroPRSMH) group at the University of Minnesota for helpful discussions.

Endnote

[1]Construct validity has also been used in recent publications as a means of testing the validity of a specific manipulation [43]. This is

usually used as a justification for the idea that one is testing the construct of a gene variant, and thus asking whether an animal model is 'valid'. It is our contention that this is a misuse of the term 'construct validity' (and a misuse of models in general). Models are a tool through which one can explore effects and consequences. Our contention is that validity is a form of measurement, thus the question should not be one of whether a model is valid or not, but rather, how valid the model is under different measurements. Moreover, the validity of a model depends on the specific questions it is addressing and the specific context in which it is applied. Asking a question of the consequences of a gene variant is a question about underlying physiology, and thus a question of 'mechanistic validity'.

References

1. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, de Vet HCW. 2010 The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J. Clin. Epidemiol.* **63**, 737–745. (doi:10.1016/j.jclinepi.2010.02.006)
2. Monteggia LM, Heimer H, Nestler EJ. 2018 Meeting report: can we make animal models of human mental illness? *Biol. Psychiatry* **84**, 542–545. (doi:10.1016/j.biopsych.2018.02.010)
3. Coolidge FL, Segal DL. 2010 Validity. In *The Corsini Encyclopedia of Psychology* (eds I Weiner, WE Craghead), pp. 1448–1449. New York, NY: Wiley. (doi:10.1002/9780470479216.corpsy1019)
4. Keifer J, Summers CH. 2016 Putting the 'biology' back into 'neurobiology': the strength of diversity in animal model systems for neuroscience research. *Front. Syst. Neurosci.* **10**, 69. (doi:10.3389/fnsys.2016.00069)
5. Blanchard DC, Summers CH, Blanchard RJ. 2013 The role of behavior in translational models for psychopathology: functionality and dysfunctional behaviors. *Neurosci. Biobehav. Rev.* **37**, 1567–1577. (doi:10.1016/j.neubiorev.2013.06.008)
6. Young JW, Amitai N, Geyer MA. 2012 Behavioral animal models to assess pro-cognitive treatments for schizophrenia. In *Novel antischizophrenia treatments* (eds MA Geyer, G Gross), pp. 39–79. Berlin, Germany: Springer.
7. Anastasi A. 1950 The concept of validity in the interpretation of test scores. *Educ. Psychol. Meas.* **10**, 67–78. (doi:10.1177/001316445001000105)
8. Cureton EE. 1965 Reliability and validity: basic assumptions and experimental designs. *Educ. Psychol. Meas.* **25**, 327–346. (doi:10.1177/001316446502500204)
9. O'Connor EC, Chapman K, Butler P, Mead AN. 2010 The predictive validity of the rat self-administration model for abuse liability. *Neurosci. Biobehav. Rev.* **35**, 912–938. (doi:10.1016/j.neubiorev.2010.10.012)
10. Fleming EG, Fleming CW. 1929 The validity of the Matthews' revision of the Woodworth personal data questionnaire. *J. Abnorm. Soc. Psychol.* **23**, 500–506. (doi:10.1037/h0075316)
11. Belzung C, Lemoine M. 2011 Criteria of validity for animal models of psychiatric disorders: focus on anxiety disorders and depression. *Biol. Mood Anxiety Disord.* **1**, 9. (doi:10.1186/2045-5380-1-9)
12. MacCorquodale K, Meehl PE. 1948 On a distinction between hypothetical constructs and intervening variables. *Psychol. Rev.* **55**, 95–107. (doi:10.1037/h0056029)
13. Cronbach LJ, Meehl PE. 1955 Construct validity in psychological tests. *Psychol. Bull.* **52**, 281–302. (doi:10.1037/h0040957)
14. Loevinger J. 1957 Objective tests as instruments of psychological theory. *Psychol. Rep.* **3**, 635–694. (doi:10.2466/pr0.1957.3.3.635)
15. Campbell DT, Fiske DW. 1959 Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol. Bull.* **56**, 81–105. (doi:10.1037/h0046016)
16. Strauss ME, Smith GT. 2009 Construct validity: advances in theory and methodology. *Annu. Rev. Clin. Psychol.* **5**, 1–25. (doi:10.1146/annurev.clinpsy.032408.153639)
17. Gallery ME, Hofmeister A. 1978 A method for assessing the treatment validity of tests in special education. *Except. Child.* **25**, 105–113. (doi:10.1080/0156655780250203)
18. Redish AD. 2013 *The mind within the brain: how we make decisions and how those decisions go wrong*. New York, NY: Oxford University Press.
19. Redish AD, Kazinka R, Herman AB. 2019 Taking an engineer's view: implications of network analysis for computational psychiatry. *Behav. Brain Sci.* **42**, e24. (doi:10.1017/S0140525X18001152)
20. Tonigan JS, Rice SL. 2010 Is it beneficial to have an alcoholics anonymous sponsor? *Psychol. Addict. Behav.* **24**, 397–403. (doi:10.1037/a0019013)
21. Harlow HF. 1971 *Learning to love*. San Francisco, CA: Albion Publishing Co.
22. Bowlby J. 1973 *Separation: anxiety and anger. Attachment and loss: volume II*. London, UK: The Hogarth Press and the Institute of Psycho-Aalysis.
23. Blum D. 2002 *Love at Goon Park: Harry Harlow and the science of affection*. New York, NY: Perseus Books.
24. Stockdale SE, Wells KB, Tang L, Belin TR, Zhang L, Sherbourne CD. 2007 The importance of social context: neighborhood stressors, stress-buffering mechanisms, and alcohol, drug, and mental health disorders. *Soc. Sci. Med.* **65**, 1867–1881. (doi:10.1016/j.socscimed.2007.05.045)
25. Whitaker LR, Degoulet M, Morikawa H. 2013 Social deprivation enhances VTA synaptic plasticity and drug-induced contextual learning. *Neuron* **77**, 335–345. (doi:10.1016/j.neuron.2012.11.022)
26. Venniro M, Zhang M, Caprioli D, Hoots JK, Golden SA, Heins C, Morales M, Epstein DH, Shaham Y. 2018 Volitional social interaction prevents drug addiction in rat models. *Nat. Neurosci.* **21**, 1520–1529. (doi:10.1038/s41593-018-0246-6)
27. Myers K, Davis M. 2007 Mechanisms of fear extinction. *Mol. Psychiatry* **12**, 120–150. (doi:10.1038/sj.mp.4001939)
28. Shansky RM, Woolley CS. 2016 Considering sex as a biological variable will be valuable for neuroscience research. *J. Neurosci.* **36**, 11 817–11 822. (doi:10.1523/JNEUROSCI.1390-16.2016)
29. Ly V, Huys QJM, Stins JF, Roelofs K, Cools R. 2014 Individual differences in bodily freezing predict emotional biases in decision making. *Front. Behav. Neurosci.* **8**, 237. (doi:10.3389/fnbeh.2014.00237)
30. Koob GF, Le Moal M. 2006 *Neurobiology of addiction*. Amsterdam, The Netherlands: Elsevier Academic Press.
31. Casper RC, Sullivan EL, Tecott L. 2008 Relevance of animal models to human eating disorders and obesity. *Psychopharmacology* (*Berl.*) **199**, 313–329. (doi:10.1007/s00213-008-1102-2)
32. Stice E. 2002 Risk and maintenance factors for eating pathology: a meta-analytic review. *Psychol. Bull.* **128**, 825–848. (doi:10.1037/0033-2909.128.5.825)
33. Smeyne RJ, Jackson-Lewis V. 2005 The MPTP model of Parkinson's disease. *Brain Res. Mol. Brain Res.* **134**, 57–66. (doi:10.1016/j.molbrainres.2004.09.017)
34. Langston JW. 1996 The etiology of Parkinson's disease with emphasis on the MPTP story. *Neurology* **47**, S153–S160. (doi:10.1212/WNL.47.6_Suppl_3.153S)
35. Dorval AD, Grill WM. 2014 Deep brain stimulation of the subthalamic nucleus reestablishes neuronal information transmission in the 6-OHDA rat model of parkinsonism. *J. Neurophysiol.* **111**, 1949–1959. (doi:10.1152/jn.00713.2013)
36. LeDoux J. 2007 The amygdala. *Curr. Biol.* **17**, R868–R874. (doi:10.1016/j.cub.2007.08.005)
37. Kolb B. 1990 Prefrontal cortex. In *The cerebral cortex of the rat* (eds B Kolb, RC Tees), pp. 437–458. Cambridge, MA: MIT Press.
38. Uylings HBM, Groenewegen HJ, Kolb B. 2003 Do rats have a prefrontal cortex? *Behav. Brain Res.* **146**, 3–17. (doi:10.1016/j.bbr.2003.09.028)
39. Wise SP. 2008 Forward frontal fields: phylogeny and fundamental function. *Trends Neurosci.* **31**, 599–608. (doi:10.1016/j.tins.2008.08.008)
40. Heilbronner SR, Rodriguez-Romaguera J, Quirk GJ, Groenewegen HJ, Haber SN. 2016 Circuit based cortico-striatal homologies between rat and primate. *Biol. Psychiatry* **80**, 509–521. (doi:10.1016/j.biopsych.2016.05.012)
41. Widge AS, Heilbronner SR, Hayden BY. 2019 Prefrontal cortex and cognitive control: new insights

from human electrophysiology. *F1000Res* **8**, 1696. (doi:10.12688/f1000research.20044.1)

42. Striedter GF. 2006 Précis of principles of brain evolution. *Behav. Brain Sci.* **29**, 1–12; discussion 12–36. (doi:10.1017/S014052 5X06009010)
43. Nestler EJ, Hyman SE. 2010 Animal models of neuropsychiatric disorders. *Nat. Neurosci.* **13**, 1161–1169. (doi:10.1038/nn.2647)
44. Luria AR. 1976 *The neuropsychology of memory*. New York. NY: John Wiley and Sons.
45. Cohen NJ, Eichenbaum H. 1993 *Memory, amnesia, and the hippocampal system*. Cambridge, MA: MIT Press.
46. Squire LR. 1987 *Memory and brain*. New York, NY: Oxford University Press.
47. Murray EA, Mishkin M. 1998 Object recognition and location memory in monkeys with excitotoxic lesions of the amygdala and hippocampus. *J. Neurosci.* **18**, 6568–6582. (doi:10.1523/JNEUROSCI.18-16-06568.1998)
48. Rudebeck PH, Saunders RC, Prescott AT, Chau LS, Murray EA. 2013 Prefrontal mechanisms of behavioral flexibility, emotion regulation and value updating. *Nat. Neurosci.* **16**, 1140–1145. (doi:10.1038/nn.3440)
49. Morris RGM, Garrud P, Rawlins JNP, O'Keefe J. 1982 Place navigation impaired in rats with hippocampal lesions. *Nature* **297**, 681–683. (doi:10.1038/297681a0)
50. Redish AD. 1999 *Beyond the cognitive map: from place cells to episodic memory*. Cambridge, MA: MIT Press.
51. Goldman-Rakic PS. 1995 Cellular basis of working memory. *Neuron* **14**, 477–485. (doi:10.1016/0896-6273(95)90304-6)
52. Fuster JM. 2008 *The prefrontal cortex*. 4th edn. Amsterdam, The Netherlands: Elsevier.
53. Redish AD. 2001 The hippocampal debate: are we asking the right questions? *Behav. Brain Res.* **127**, 81–98. (doi:10.1016/S0166-4328(01)00356-4)
54. Marr D. 1982 *Vision*. New York, NY: W. H. Freeman and Co.
55. Levenstein D *et al*. 2020 On the role of theory and modeling in neuroscience. *arXiv* 2003.13825 [q-bio.NC].
56. Baddeley A. 1992 Working memory. *Science* **255**, 556–559. (doi:10.1126/science.1736359)
57. Fuster JM. 1990 Prefrontal cortex and the bridging of temporal gaps in the perception-action cycle. *Ann. NY Acad. Sci.* **608**, 318–329; discussion 330–336. (doi:10.1111/j.1749-6632.1990.tb48901.x)
58. Churchland P, Sejnowski TJ. 1994 *The computational brain*. Cambridge, MA: MIT Press.
59. Milner D, Goodale M. 2006 *The visual brain in action*. New York, NY: Oxford University Press.
60. O'Keefe J, Nadel L. 1978 *The hippocampus as a cognitive Map*. Oxford, UK: Clarendon Press.
61. Gilbert DT, Wilson TD. 2007 Prospection: experiencing the future. *Science* **317**, 1351–1354. (doi:10.1126/science.1144161)
62. Buckner RL, Carroll DC. 2007 Self-projection and the brain. *Trends Cogn. Sci.* **11**, 49–57. (doi:10.1016/j.tics.2006.11.004)
63. Redish AD. 2016 Vicarious trial and error. *Nat. Rev. Neurosci.* **17**, 147–159. (doi:10.1038/nrn.2015.30)
64. Johnson A, van der Meer MAA, Redish AD. 2007 Integrating hippocampus and striatum in decision-making. *Curr. Opin. Neurobiol.* **17**, 692–697. (doi:10.1016/j.conb.2008.01.003)
65. Daw ND, Niv Y, Dayan P. 2005 Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.* **8**, 1704–1711. (doi:10.1038/nn1560)
66. Niv Y, Joel D, Dayan P. 2006 A normative perspective on motivation. *Trends Cogn. Sci.* **10**, 375–381. (doi:10.1016/j.tics.2006.06.010)
67. Johnson A, Redish AD. 2007 Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point. *J. Neurosci.* **27**, 12 176–12 189. (doi:10.1523/JNEUROSCI.3761-07.2007)
68. Kay K, Chung JE, Sosa M, Schor JS, Karlsson MP, Larkin MC, Liu DF, Frank LM. 2020 Constant sub-second cycling between representations of possible futures in the hippocampus. *Cell* **180**, 552–567.e25. (doi:10.1016/j.cell.2020.01.014)
69. Schmidt B, Duin AA, Redish AD. 2019 Disrupting the medial prefrontal cortex alters hippocampal sequences during deliberative decision making. *J. Neurophysiol.* **121**, 1981–2000. (doi:10.1152/jn.00793.2018)
70. Ito HT, Zhang S-J, Witter MP, Moser EI, Moser M-B. 2015 A prefrontal-thalamo-hippocampal circuit for goal-directed spatial navigation. *Nature* **522**, 50–55. (doi:10.1038/nature14396)
71. Wang JX, Cohen NJ, Voss JL. 2015 Covert rapid action-memory simulation (CRAMS): a hypothesis of hippocampal–prefrontal interactions for adaptive behavior. *Neurobiol. Learn. Mem.* **117**, 22–33. (doi:10.1016/j.nlm.2014.04.003)
72. van der Meer MAA, Johnson A, Schmitzer-Torbert NC, Redish AD. 2010 Triple dissociation of information processing in dorsal striatum, ventral striatum, and hippocampus on a learned spatial decision task. *Neuron* **67**, 25–32. (doi:10.1016/j.neuron.2010.06.023)
73. Abraham L, Potegal M, Miller S. 1983 Evidence for caudate nucleus involvement in an egocentric spatial task: return from passive transport. *Physiol. Psychol.* **11**, 11–17. (doi:10.3758/BF03326764)
74. Yin HH, Knowlton B, Balleine BW. 2004 Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. *Eur. J. Neurosci.* **19**, 181–189. (doi:10.1111/j.1460-9568.2004.03095.x)
75. Roitman JD, Shadlen MN. 2002 Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *J. Neurosci.* **22**, 9475–9489. (doi:10.1523/JNEUROSCI.22-21-09475.2002)
76. Palmer J, Huk AC, Shadlen MN. 2005 The effect of stimulus strength on the speed and accuracy of a perceptual decision. *J. Vis.* **5**, 376–404. (doi:10.1167/5.5.1)
77. Britten KH, Shadlen MN, Newsome WT, Movshon JA. 1992 The analysis of visual motion: a comparison of neuronal and psychophysical performance. *J. Neurosci.* **12**, 4745–4765. (doi:10.1523/JNEUROSCI.12-12-04745.1992)
78. Brunton BW, Botvinick MM, Brody CD. 2013 Rats and humans can optimally accumulate evidence for decision-making. *Science* **340**, 95–98. (doi:10.1126/science.1233912)
79. Sanders JI, Kepecs A. 2012 Choice ball: a response interface for two-choice psychometric discrimination in head-fixed mice. *J. Neurophysiol.* **108**, 3416–3423. (doi:10.1152/jn.00669.2012)
80. Pinto L, Koay SA, Engelhard B, Yoon AM, Deverett B, Thiberge SY, Witten IB, Tank DW, Brody CD. 2018 An accumulation-of-evidence task using visual pulses for mice navigating in virtual reality. *Front. Behav. Neurosci.* **12**, 36. (doi:10.3389/fnbeh.2018.00036)
81. Constantinople CM, Piet AT, Brody CD. 2019 An analysis of decision under risk in rats. *Curr. Biol.* **29**, 2066–2074.e5. (doi:10.1016/j.cub.2019.05.013)
82. Mazurek ME, Roitman JD, Ditterich J, Shadlen MN. 2003 A role for neural integrators in perceptual decision making. *Cereb. Cortex* **13**, 1257–1269. (doi:10.1093/cercor/bhg097)
83. Ding L, Gold JI. 2010 Caudate encodes multiple computations for perceptual decisions. *J. Neurosci.* **30**, 15 747–15 759. (doi:10.1523/JNEUROSCI.2894-10.2010)
84. Hanks TD, Kopec CD, Brunton BW, Duan CA, Erlich JC, Brody CD. 2015 Distinct relationships of parietal and prefrontal cortices to evidence accumulation. *Nature* **520**, 220–223. (doi:10.1038/nature14066)
85. Erlich JC, Brunton BW, Duan CA, Hanks TD, Brody CD. 2015 Distinct effects of prefrontal and parietal cortex inactivations on an accumulation of evidence task in the rat. *Elife* **4**, e05457. (doi:10.7554/elife.05457)
86. Katz LN, Yates JL, Pillow JW, Huk AC. 2016 Dissociated functional significance of decision-related activity in the primate dorsal stream. *Nature* **535**, 285–288. (doi:10.1038/nature18617)
87. Gold JI, Shadlen MN. 2001 Neural computations that underlie decisions about sensory stimuli. *Trends Cogn. Sci.* **5**, 10–16. (doi:10.1016/S1364-6613(00)01567-9)
88. Papale AE, Stott JJ, Powell NJ, Regier PS, Redish AD. 2012 Interactions between deliberation and delay discounting in rats. *Cogn. Affect. Behav. Neurosci.* **12**, 513–526. (doi:10.3758/s13415-012-0097-7)
89. Cardinal RN, Daw N, Robbins T, Everitt BJ. 2002 Local analysis of behaviour in the adjusting-delay task for choice of delayed reinforcement. *Neural Netw.* **15**, 617–634. (doi:10.1016/S0893-6080(02)00053-9)
90. McDonald RJ, White NM. 1994 Parallel information processing in the water maze: evidence for independent memory systems involving dorsal striatum and hippocampus. *Behav. Neural. Biol.* **61**, 260–270. (doi:10.1016/S0163-1047(05)80009-3)

91. Sweis BM, Thomas MJ, Redish AD. 2018 Mice learn to avoid regret. *PLoS Biol.* **16**, e2005853. (doi:10.1371/journal.pbio.2005853)
92. Steiner AP, Redish AD. 2014 Behavioral and neurophysiological correlates of regret in rat decision-making on a neuroeconomic task. *Nat. Neurosci.* **17**, 995–1002. (doi:10.1038/nn.3740)
93. Sweis BM, Abram SV, Schmidt BJ, Seeland KD, MacDonald AW, Thomas MJ, Redish AD. 2018 Sensitivity to 'sunk costs' in mice, rats, and humans. *Science* **361**, 178–181. (doi:10.1126/science.aar8644)
94. Kalenscher T, Wingerden MV. 2011 Why we should use animals to study economic decision making? A perspective. *Front. Neurosci.* **5**, 82. (doi:10.3389/fnins.2011.00082)
95. Fiddick L, Cosmides L, Tooby J. 2000 No interpretation without representation: the role of domain-specific representations and inferences in the Wason selection task. *Cognition* **77**, 1–79. (doi:10.1016/S0010-0277(00)00085-8)
96. Tolman EC, Ritchie BF, Kalish D. 1946 Studies in spatial learning. II. Place learning versus response learning. *J. Exp. Psychol.* **36**, 221–229. (doi:10.1037/h0060262)
97. Hull CL. 1943 *Principles of behavior*. New York, NY: Appleton-Century-Crofts.
98. Barnes CA, Nadel L, Honig WK. 1980 Spatial memory deficit in senescent rats. *Can. J. Psychol.* **34**, 29–39. (doi:10.1037/h0081022)
99. Packard MG, McGaugh JL. 1996 Inactivation of hippocampus or caudate nucleus with lidocaine differentially affects expression of place and response learning. *Neurobiol. Learn. Mem.* **65**, 65–72. (doi:10.1006/nlme.1996.0007)
100. Tolman EC. 1939 Prediction of vicarious trial and error by means of the schematic sowbug. *Psychol. Rev.* **46**, 318–336. (doi:10.1037/h0057054)
101. Tolman EC. 1948 Cognitive maps in rats and men. *Psychol. Rev.* **55**, 189–208. (doi:10.1037/h0061626)
102. van der Meer MAA, Kurth-Nelson Z, Redish AD. 2012 Information processing in decision-making systems. *Neuroscientist* **18**, 342–359. (doi:10.1177/1073858411435128)
103. Redish AD, Jensen S, Johnson A, Kurth-Nelson Z. 2007 Reconciling reinforcement learning models with behavioral extinction and renewal: implications for addiction, relapse, and problem gambling. *Psychol. Rev.* **114**, 784–805. (doi:10.1037/0033-295X.114.3.784)
104. Yin HH, Knowlton BJ. 2004 Contributions of striatal subregions to place and response learning. *Learn. Mem.* **11**, 459–463. (doi:10.1101/lm.81004)
105. Regier PS, Amemiya S, Redish AD. 2015 Hippocampus and subregions of the dorsal striatum respond differently to a behavioral strategy change on a spatial navigation task. *J. Neurophysiol.* **114**, 1399–1416. (doi:10.1152/jn.00189.2015)
106. Hasz BM, Redish AD. 2020 Dorsomedial prefrontal cortex and hippocampus represent strategic context even while simultaneously changing representation throughout a task session. *Neurobiol. Learn. Mem.* **171**, 107215. (doi:10.1016/j.nlm.2020.107215)
107. Powell NJ, Redish AD. 2016 Representational changes of latent strategies in rat medial prefrontal cortex precede changes in behaviour. *Nat. Commun.* **7**, 12830. (doi:10.1038/ncomms12830)
108. Rand MK, Hikosaka O, Miyachi S, Lu X, Miyashita K. 1998 Characteristics of a long-term procedural skill in the monkey. *Exp. Brain Res.* **118**, 293–297. (doi:10.1007/s002210050284)
109. Rand MK, Hikosaka O, Miyachi S, Lu X, Nakamura K, Kitaguchi K, Shimo Y. 2000 Characteristics of sequential movements during early learning period in monkeys. *Exp. Brain Res.* **131**, 293–304. (doi:10.1007/s002219900283)
110. Hikosaka O, Nakahara H, Rand MK, Sakai K, Lu X, Nakamura K, Miyachi S, Doya K. 1999 Parallel neural networks for learning sequential procedures. *Trends Neurosci.* **22**, 464–471. (doi:10.1016/S0166-2236(99)01439-3)
111. Desrochers TM, Amemori K-I, Graybiel AM. 2015 Habit learning by naive macaques is marked by response sharpening of striatal neurons representing the cost and outcome of acquired action sequences. *Neuron* **87**, 853–868. (doi:10.1016/j.neuron.2015.07.019)
112. Graybiel AM. 1998 The basal ganglia and chunking of action repertoires. *Neurobiol. Learn. Mem.* **70**, 119–136. (doi:10.1006/nlme.1998.3843)
113. Knopman D, Nissen MJ. 1991 Procedural learning is impaired in Huntington's disease: evidence from the serial reaction time task. *Neuropsychologia* **29**, 245–254. (doi:10.1016/0028-3932(91)90085-M)
114. Jackson GM, Jackson SR, Harrison J, Henderson L, Kennard C. 1995 Serial reaction time learning and Parkinson's disease: evidence for a procedural learning deficit. *Neuropsychologia* **33**, 577–593. (doi:10.1016/0028-3932(95)00010-Z)
115. Jackson GM, Jackson SR. 1995 Do measures of explicit learning actually measure what is being learnt in the serial reaction time task? A critique of current methods. *Psyche* **2**, 20.
116. Steiner A, Redish AD. 2012 Orbitofrontal cortical ensembles during deliberation and learning on a spatial decision-making task. *Front. Decis. Neurosci.* **6**, 131.
117. Blumenthal A, Steiner A, Seeland KD, Redish AD. 2011 Effects of pharmacological manipulations of NMDA-receptors on deliberation in the multiple-T task. *Neurobiol. Learn. Mem.* **95**, 376–384. (doi:10.1016/j.nlm.2011.01.011)
118. Smith AC, Frank LM, Wirth S, Yanike M, Hu D, Kubota Y, Graybiel AM, Suzuki WA, Brown EN. 2004 Dynamic analysis of learning in behavioral experiments. *J. Neurosci.* **24**, 447–461. (doi:10.1523/JNEUROSCI.2908-03.2004)
119. Cohen NJ, Squire LR. 1980 Preserved learning and retention of pattern-analyzing skill in amnesia: dissociation of knowing how and knowing that. *Science* **210**, 207–210. (doi:10.1126/science.7414331)
120. Gardner RS, Uttaro MR, Fleming SE, Suarez DF, Ascoli GA, Dumas TC. 2013 A secondary working memory challenge preserves primary place strategies despite over-training. *Learn. Mem.* **20**, 648–656. (doi:10.1101/lm.031336.113)
121. Schmidt BJ, Papale AE, Redish AD, Markus EJ. 2013 Conflict between place and response navigation strategies: effects on vicarious trial and error (VTE) behaviors. *Learn. Mem.* **20**, 130–138. (doi:10.1101/lm.028753.112)
122. Johnson A, Fenton AA, Kentros C, Redish AD. 2009 Looking for cognition in the structure in the noise. *Trends Cogn. Sci.* **13**, 55–64. (doi:10.1016/j.tics.2008.11.005)
123. Amemiya S, Redish AD. 2016 Manipulating decisiveness in decision making—effects of clonidine on hippocampal search strategies. *J. Neurosci.* **36**, 814–827. (doi:10.1523/JNEUROSCI.2595-15.2016)
124. Wikenheiser AM, Redish AD. 2015 Hippocampal theta sequences reflect current goals. *Nat. Neurosci.* **18**, 289–294. (doi:10.1038/nn.3909)
125. Papale AE, Zielinski MC, Frank LM, Jadhav SP, Redish AD. 2016 Interplay between hippocampal sharp-wave-ripple events and vicarious trial and error behaviors in decision making. *Neuron* **92**, 975–982. (doi:10.1016/j.neuron.2016.10.028)
126. Allen TA, Salz DM, McKenzie S, Fortin NJ. 2016 Nonspatial sequence coding in CA1 neurons. *J. Neurosci.* **36**, 1547–1563. (doi:10.1523/JNEUROSCI.2874-15.2016)
127. van der Meer MAA, Redish AD. 2009 Covert expectation-of-reward in rat ventral striatum at decision points. *Front. Integr. Neurosci.* **3**, 1–15. (doi:10.3389/neuro.07.001.2009)
128. van der Meer MAA, Redish AD. 2011 Theta phase precession in rat ventral striatum links place and reward information. *J. Neurosci.* **31**, 2843–2854. (doi:10.1523/JNEUROSCI.4869-10.2011)
129. Jog MS, Kubota Y, Connolly CI, Hillegaart V, Graybiel AM. 1999 Building neural representations of habits. *Science* **286**, 1746–1749.
130. Barnes TD, Kubota Y, Hu D, Jin DZ, Graybiel AM. 2005 Activity of striatal neurons reflects dynamic encoding and recoding of procedural memories. *Nature* **437**, 1158–1161. (doi:10.1038/nature04053)
131. Smith KS, Graybiel AM. 2013 A dual operator view of habitual behavior reflecting cortical and striatal dynamics. *Neuron* **79**, 361–374. (doi:10.1016/j.neuron.2013.05.038)
132. Schmitzer-Torbert NC, Redish AD. 2004 Neuronal activity in the rodent dorsal striatum in sequential navigation: separation of spatial and reward responses on the multiple-T task. *J. Neurophysiol.* **91**, 2259–2272. (doi:10.1152/jn.00687.2003)
133. Schmitzer-Torbert NC, Redish AD. 2008 Task-dependent encoding of space and events by striatal neurons is dependent on neural subtype. *Neuroscience* **153**, 349–360. (doi:10.1016/j.neuroscience.2008.01.081)
134. Thorn CA, Atallah H, Howe M, Graybiel AM. 2010 Differential dynamics of activity changes in dorsolateral and dorsomedial striatal loops during learning. *Neuron* **66**, 781–795. (doi:10.1016/j.neuron.2010.04.036)

135. Yang T, Shadlen MN. 2007 Probabilistic reasoning by neurons. *Nature* **447**, 1075–1080. (doi:10.1038/nature05852)
136. Kepecs A, Uchida N, Zariwala HA, Mainen ZF. 2008 Neural correlates, computation and behavioural impact of decision confidence. *Nature* **455**, 227–231. (doi:10.1038/nature07200)
137. Ott T, Masset P, Kepecs A. 2018 The neurobiology of confidence: from beliefs to neurons. *Cold Spring Harb. Symp. Quant. Biol.* **83**, 9–16. (doi:10.1101/sqb.2018.83.038794)
138. Link SW, Heath RA. 1975 A sequential theory of psychological discrimination. *Psychometrika* **40**, 77–105. (doi:10.1007/BF02291481)
139. Ratcliff R. 1988 Continuous versus discrete information processing modeling accumulation of partial information. *Psychol. Rev.* **95**, 238–255. (doi:10.1037/0033-295X.95.2.238)
140. Ratcliff R, Rouder JN. 1998 Modeling response times for two-choice decisions. *Psychol. Sci.* **9**, 347–356. (doi:10.1111/1467-9280.00067)
141. Glaze CM, Kable JW, Gold JI. 2015 Normative evidence accumulation in unpredictable environments. *Elife* **4**, e08825. (doi:10.7554/eLife.08825)
142. Ratcliff R, Smith PL, Brown SD, McKoon G. 2016 Diffusion decision model: current issues and history. *Trends Cogn. Sci.* **20**, 260–281. (doi:10.1016/j.tics.2016.01.007)
143. Downing CJ, Movshon JA. 1989 Spatial and temporal summation in the detection of motion in stochastic random dot displays. *Invest. Ophthalmol. Vis. Sci.* **30**, 72.
144. Bogacz R, Brown E, Moehlis J, Holmes P, Cohen JD. 2006 The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychol. Rev.* **113**, 700–765. (doi:10.1037/0033-295X.113.4.700)
145. Forstmann BU, Ratcliff R, Wagenmakers E-J. 2016 Sequential sampling models in cognitive neuroscience: advantages, applications, and extensions. *Annu. Rev. Psychol.* **67**, 641–666. (doi:10.1146/annurev-psych-122414-033645)
146. Cisek P, Puskas GA, El-Murr S. 2009 Decisions in changing conditions: the urgency-gating model. *J. Neurosci.* **29**, 11 560–11 571. (doi:10.1523/JNEUROSCI.1844-09.2009)
147. Hanks T, Kiani R, Shadlen MN. 2014 A neural mechanism of speed-accuracy tradeoff in macaque area LIP. *Elife* **3**, e02260. (doi:10.7554/eLife.02260)
148. Morita T, Kang H, Wolfe J, Jadhav SP, Feldman DE. 2011 Psychometric curve and behavioral strategies for whisker-based texture discrimination in rats. *PLoS ONE* **6**, e20437. (doi:10.1371/journal.pone.0020437)
149. Raposo D, Sheppard JP, Schrater PR, Churchland AK. 2012 Multisensory decision-making in rats and humans. *J. Neurosci.* **32**, 3726–3735. (doi:10.1523/JNEUROSCI.4998-11.2012)
150. Brody CD, Hanks TD. 2016 Neural underpinnings of the evidence accumulator. *Curr. Opin. Neurobiol.* **37**, 149–157. (doi:10.1016/j.conb.2016.01.003)
151. Davis M, Falls WA, Campeau S, Kim M. 1993 Fear-potentiated startle: a neural and pharmacological analysis. *Behav. Brain Res.* **58**, 175–198. (doi:10.1016/0166-4328(93)90102-V)
152. LeDoux JE. 2015 *Anxious: using the brain to understand and treat fear and anxiety*. New York, NY: Penguin.
153. Pellman BA, Kim JJ. 2016 What can ethobehavioral studies tell us about the brain's fear system? *Trends Neurosci.* **39**, 420–431. (doi:10.1016/j.tins.2016.04.001)
154. Kim JJ, Jung MW. 2018 Fear paradigms: the times they are a-changin'. *Curr. Opin. Behav. Sci.* **24**, 38–43. (doi:10.1016/j.cobeha.2018.02.007)
155. Rogan MT, Stäubli UV, LeDoux JE. 1997 Fear conditioning induces associative long-term potentiation in the amygdala. *Nature* **390**, 604–607. (doi:10.1038/37601)
156. Nabavi S, Fox R, Proulx CD, Lin JY, Tsien RY, Malinow R. 2014 Engineering a memory with LTD and LTP. *Nature* **511**, 348–352. (doi:10.1038/nature13294)
157. Mobbs D, Marchant JL, Hassabis D, Seymour B, Tan G, Gray M, Petrovic P, Dolan RJ, Frith CD. 2009 From threat to fear: the neural organization of defensive fear systems in humans. *J. Neurosci.* **29**, 12 236–12 243. (doi:10.1523/JNEUROSCI.2378-09.2009)
158. Mobbs D, Hagan CC, Dalgleish T, Silston B, Prévost C. 2015 The ecology of human fear: survival optimization and the nervous system. *Front. Neurosci.* **9**, 55. (doi:10.3389/fnins.2015.00055)
159. Wu JQ, Szpunar KK, Godovich SA, Schacter DL, Hofmann SG. 2015 Episodic future thinking in generalized anxiety disorder. *J. Anxiety Disord.* **36**, 1–8. (doi:10.1016/j.janxdis.2015.09.005)
160. Mobbs D, Headley DB, Ding W, Dayan P. 2020 Space, time, and fear: survival computations along defensive circuits. *Trends Cogn. Sci.* **24**, 228–241. (doi:10.1016/j.tics.2019.12.016)
161. Beck AT, Emery G, Greenberg RL. 2005 *Anxiety disorders and phobias: a cognitive perspective*. New York, NY: Basic Books.
162. McNally GP. 2021 Motivational competition and the paraventricular thalamus. *Neurosci. Biobehav. Rev.* **125**, 193–207. (doi:10.1016/j.neubiorev.2021.02.021)
163. Fanselow MS, Lester LS. 1988 A functional behavioristic approach to aversively motivated behavior: predatory imminence as a determinant of the topography of defensive behavior. In *Evolution and learning* (ed. RC Bolles), pp. 185–212. Hillsdale, NJ: Lawrence Erlbaum Associates.
164. Mobbs D, Kim JJ. 2015 Neuroethological studies of fear, anxiety, and risky decision-making in rodents and humans. *Curr. Opin. Behav. Sci.* **5**, 8–15. (doi:10.1016/j.cobeha.2015.06.005)
165. Choi J-S, Kim JJ. 2010 Amygdala regulates risk of predation in rats foraging in a dynamic fear environment. *Proc. Natl Acad. Sci. USA* **107**, 21 773–21 777. (doi:10.1073/pnas.1010079108)
166. Walters CJ, Jubran J, Sheehan A, Erickson MT, Redish AD. 2019 Avoid-approach conflict behaviors differentially affected by anxiolytics: implications for a computational model of risky decision-making. *Psychopharmacology* (*Berl.*) **236**, 2513–2525. (doi:10.1007/s00213-019-05197-0)
167. Amir A, Lee S-C, Headley DB, Herzallah MM, Pare D. 2015 Amygdala signaling during foraging in a hazardous environment. *J. Neurosci.* **35**, 12 994–13 005. (doi:10.1523/JNEUROSCI.0407-15.2015)
168. Morriss J, Christakou A, van Reekum CM. 2016 Nothing is safe: intolerance of uncertainty is associated with compromised fear extinction learning. *Biol. Psychol.* **121**, 187–193. (doi:10.1016/j.biopsycho.2016.05.001)
169. Koenig S, Uengoer M, Lachnit H. 2017 Attentional bias for uncertain cues of shock in human fear conditioning: evidence for attentional learning theory. *Front. Hum. Neurosci.* **11**, 266. (doi:10.3389/fnhum.2017.00266)
170. Kepecs A, Mensh BD. 2015 Emotor control: computations underlying bodily resource allocation, emotions, and confidence. *Dialogues Clin. Neurosci.* **17**, 391–401. (doi:10.31887/DCNS.2015.17.4/akepecs)
171. Molewijk HE, der Poel AM van, Oliver B. 1995 The ambivalent behaviour 'stretched approach posture' in the rat as a paradigm to characterize anxiolytic drugs. *Psychopharmacology* (*Berl.*) **121**, 81–90. (doi:10.1007/BF02245594)
172. Grewal SS, Shepherd JK, Bill DJ, Fletcher A, Dourish CT. 1997 Behavioural and pharmacological characterisation of the canopy stretched attend posture test as a model of anxiety in mice and rats. *Psychopharmacology* (*Berl.*) **133**, 29–38. (doi:10.1007/s002130050367)
173. Gray J, McNaughton N. 2000 *The neuropsychology of anxiety*. New York, NY: Oxford University Press.
174. Wu C-T, Haggerty D, Kemere C, Ji D. 2017 Hippocampal awake replay in fear memory retrieval. *Nat. Neurosci.* **20**, 571–580. (doi:10.1038/nn.4507)
175. Heller AS, Bagot RC. 2020 Is hippocampal replay a mechanism for anxiety and depression? *JAMA Psychiatry* **77**, 431–432. (doi:10.1001/jamapsychiatry.2019.4788)
176. Walters CJ. 2021 Neural computations underpinning anxiety in health and disease. PhD thesis, University of Minnesota, MN.
177. Abram SV, Breton Y-A, Schmidt B, Redish AD, MacDonald III AW. 2016 The web-surf task: a translational model of human decision-making. *Cogn. Affect. Behav. Neurosci.* **16**, 37–50. (doi:10.3758/s13415-015-0379-y)
178. Kazinka R, MacDonald III AW, Redish AD. 2021 Sensitivity to sunk costs depends on attention to the delay. *Front. Psychol.* **12**, 604843. (doi:10.3389/fpsyg.2021.604843)
179. Huynh T, Alstatt K, Abram SV, Schmitzer-Torbert N. 2021 Vicarious trial-and-error is enhanced during deliberation in human virtual navigation in a translational foraging task. *Front. Behav. Neurosci.* **15**, 586159. (doi:10.3389/fnbeh.2021.586159)
180. Haynos A, Abram S, Sweis B, MacDonald III AW, Redish AD, Crow S. 2019 S74. Identifying valuation disturbances in anorexia nervosa using a

translational decision-making paradigm. *Biol. Psychiatry* **85**, S325. (doi:10.1016/j.biopsych.2019.03.825)

181. Flavell JH. 1979 Metacognition and cognitive monitoring: a new area of cognitive-developmental inquiry. *Am. Psychol.* **34**, 906–911. (doi:10.1037/0003-066x.34.10.906)
182. Metcalfe J. 2008 Evolution of metacognition. *Handb. Metamemory Mem.* **29**, 46. (doi:10.4324/9780203805503.ch3)
183. Fleming SM, Dolan RJ. 2012 The neural basis of metacognitive ability. *Phil. Trans. R. Soc. B* **367**, 1338–1349. (doi:10.1098/rstb.2011.0417)
184. Hangya B, Sanders JI, Kepecs A. 2016 A mathematical framework for statistical decision confidence. *Neural Comput.* **28**, 1840–1858. (doi:10.1162/NECO_a_00864)
185. Pouget A, Drugowitsch J, Kepecs A. 2016 Confidence and certainty: distinct probabilistic quantities for different goals. *Nat. Neurosci.* **19**, 366–374. (doi:10.1038/nn.4240)
186. Kiani R, Shadlen MN. 2009 Representation of confidence associated with a decision by neurons in the parietal cortex. *Science* **324**, 759–764. (doi:10.1126/science.1169405)
187. Middlebrooks PG, Sommer MA. 2012 Neuronal correlates of metacognition in primate frontal cortex. *Neuron* **75**, 517–530. (doi:10.1016/j.neuron.2012.05.028)
188. Komura Y, Nikkuni A, Hirashima N, Uetake T, Miyamoto A. 2013 Responses of pulvinar neurons reflect a subject's confidence in visual categorization. *Nat. Neurosci.* **16**, 749–755. (doi:10.1038/nn.3393)
189. Stolyarova A, Rakhshan M, Hart EE, O'Dell TJ, Peters MAK, Lau H, Soltani A, Izquierdo A. 2019 Contributions of anterior cingulate cortex and basolateral amygdala to decision confidence and learning under uncertainty. *Nat. Commun.* **10**, 1–4. (doi:10.1038/s41467-019-12725-1)
190. Schmack K, Bosc M, Ott T, Sturgill JF, Kepecs A. 2021 Striatal dopamine mediates hallucination-like perception in mice. *Science* **372**, eabf4740. (doi:10.1126/science.abf4740)
191. Lak A, Costa GM, Romberg E, Koulakov AA, Mainen ZF, Kepecs A. 2014 Orbitofrontal cortex is required for optimal waiting based on decision confidence. *Neuron* **84**, 190–201. (doi:10.1016/j.neuron.2014.08.039)
192. Masset P, Ott T, Lak A, Hirokawa J, Kepecs A. 2020 Behavior- and modality-general representation of confidence in orbitofrontal cortex. *Cell* **182**, 112–126; e18. (doi:10.1016/j.cell.2020.05.022)
193. Odegaard B, Grimaldi P, Cho SH, Peters MAK, Lau H, Basso MA. 2018 Superior colliculus neuronal ensemble activity signals optimal rather than subjective confidence. *Proc. Natl Acad. Sci. USA* **115**, E1588–E1597. (doi:10.1073/pnas.1711628115)
194. Goupil L, Kouider S. 2016 Behavioral and neural indices of metacognitive sensitivity in preverbal infants. *Curr. Biol.* **26**, 3038–3045. (doi:10.1016/j.cub.2016.09.004)
195. Drugowitsch J, Mendonça AG, Mainen ZF, Pouget A. 2019 Learning optimal decisions with confidence. *Proc. Natl Acad. Sci. USA* **116**, 24 872–24 880. (doi:10.1073/pnas.1906787116)
196. Meyniel F, Schlunegger D, Dehaene S. 2015 The sense of confidence during probabilistic learning: a normative account. *PLoS Comput. Biol.* **11**, e1004305. (doi:10.1371/journal.pcbi.1004305)
197. Wei Z, Wang X-J. 2015 Confidence estimation as a stochastic process in a neurodynamical system of decision making. *J. Neurophysiol.* **114**, 99–113. (doi:10.1152/jn.00793.2014)
198. Lak A *et al*. 2020 Reinforcement biases subsequent perceptual decisions when confidence is low, a widespread behavioral phenomenon. *Elife* **9**, e49834. (doi:10.7554/eLife.49834)
199. Sanders JI, Hangya B, Kepecs A. 2016 Signatures of a statistical computation in the human sense of confidence. *Neuron* **90**, 499–506. (doi:10.1016/j.neuron.2016.03.025)
200. Fleming SM, Daw ND. 2017 Self-evaluation of decision-making: a general Bayesian framework for metacognitive computation. *Psychol. Rev.* **124**, 91–114. (doi:10.1037/rev0000045)
201. Rutishauser U, Aflalo T, Rosario ER, Pouratian N, Andersen RA. 2018 Single-neuron representation of memory strength and recognition confidence in left human posterior parietal cortex. *Neuron* **97**, 209–220.e3. (doi:10.1016/j.neuron.2017.11.029)
202. Miyamoto K, Setsuie R, Osada T, Miyashita Y. 2018 Reversible silencing of the frontopolar cortex selectively impairs metacognitive judgment on non-experience in primates. *Neuron* **97**, 980–989.e6. (doi:10.1016/j.neuron.2017.12.040)
203. Hoven M, Lebreton M, Engelmann JB, Denys D, Luigjes J, van Holst RJ. 2019 Abnormalities of confidence in psychiatry: an overview and future perspectives. *Transl. Psychiatry* **9**, 268. (doi:10.1038/s41398-019-0602-7)
204. Rouault M, Seow T, Gillan CM, Fleming SM. 2018 Psychiatric symptom dimensions are associated with dissociable shifts in metacognition but not task performance. *Biol. Psychiatry* **84**, 443–451. (doi:10.1016/j.biopsych.2017.12.017)
205. Seow TXF, Gillan CM. 2020 Transdiagnostic phenotyping reveals a host of metacognitive deficits implicated in compulsivity. *Sci. Rep.* **10**, 2883. (doi:10.1038/s41598-020-59646-4)
206. Vaghi MM, Luyckx F, Sule A, Fineberg NA, Robbins TW, De Martino B. 2017 Compulsivity reveals a novel dissociation between action and confidence. *Neuron* **96**, 348–354; e4. (doi:10.1016/j.neuron.2017.09.006)
207. Banca P, Vestergaard MD, Rankov V, Baek K, Mitchell S, Lapa T, Castelo-Branco M, Voon V. 2015 Evidence accumulation in obsessive-compulsive disorder: the role of uncertainty and monetary reward on perceptual decision-making thresholds. *Neuropsychopharmacology* **40**, 1192–1202. (doi:10.1038/npp.2014.303)
208. Campbell WK, Goodie AS, Foster JD. 2004 Narcissism, confidence, and risk attitude. *J. Behav. Decis. Mak.* **17**, 297–311. (doi:10.1002/bdm.475)
209. Dunning D, Story AL. 1991 Depression, realism, and the overconfidence effect: are the sadder wiser when predicting future actions and events? *J. Pers. Soc. Psychol.* **61**, 521–532. (doi:10.1037/0022-3514.61.4.521)
210. Korn CW, Sharot T, Walter H, Heekeren HR, Dolan RJ. 2014 Depression is related to an absence of optimistically biased belief updating about future life events. *Psychol. Med.* **44**, 579–592. (doi:10.1017/S0033291713001074)
211. Higgins ST, Delaney DD, Budney AJ, Bickel WK, Hughes JR, Foerg F, Fenwick JW. 1991 A behavioral approach to achieving initial cocaine abstinence. *Am. J. Psychiatry* **148**, 1218–1224. (doi:10.1176/ajp.148.9.1218)
212. Petry NM. 2012 *Contingency management: for substance abuse treatment*. New York, NY: Routledge.
213. Hursh SR, Galuska CM, Winger G, Woods JH. 2005 The economics of drug abuse: a quantitative assessment of drug demand. *Mol. Interv.* **5**, 20–28. (doi:10.1124/mi.5.1.6)
214. Higgins ST, Heil SH, Lussier JP. 2004 Clinical implications of reinforcement as a determinant of substance use disorderse. *Annu. Rev. Psychol.* **55**, 431–461. (doi:10.1146/annurev.psych.55.090902.142033)
215. Lussier JP, Heil SH, Mongeon JA, Badger GJ, Higgins ST. 2006 A meta-analysis of voucher-based reinforcement therapy for substance use disorders. *Addiction* **101**, 192–203. (doi:10.1111/j.1360-0443.2006.01311.x)
216. Regier PS, Redish AD. 2015 Contingency management and deliberative decision-making processes. *Front. Psychiatry* **6**, 0076. (doi:10.3389/fpsyt.2015.00076)
217. Ahmed SH. 2010 Validation crisis in animal models of drug addiction: beyond non-disordered drug use toward drug addiction. *Neurosci. Biobehav. Rev.* **35**, 172–184. (doi:10.1016/j.neubiorev.2010.04.005)
218. Redish AD, Schultheiss NW, Carter EC. 2016 The computational complexity of valuation and motivational forces in decision-making processes. *Curr. Top. Behav. Neurosci.* **27**, 313–333. (doi:10.1007/7854_2015_375)
219. Schacter DL, Addis DR, Buckner RL. 2007 Remembering the past to imagine the future: the prospective brain. *Nat. Rev. Neurosci.* **8**, 657–661. (doi:10.1038/nrn2213)
220. Hassabis D, Maguire EA. 2011 The construction system in the brain. In *Predictions in the brain: using our past to generate a future* (ed. M Bar), pp. 70–82. New York, NY: Oxford University Press.
221. Bickel WK, Yi R, Landes RD, Hill PF, Baxter C. 2011 Remember the future: working memory training decreases delay discounting among stimulant addicts. *Biol. Psychiatry* **69**, 260–265. (doi:10.1016/j.biopsych.2010.08.017)

222. Radu PT, Yi R, Bickel WK, Gross JJ, McClure SM. 2011 A mechanism for reducing delay discounting by altering temporal attention. *J. Exp. Anal. Behav.* **96**, 363–385. (doi:10.1901/jeab.2011.96-363)
223. Sayegh CS, Huey SJ, Zara EJ, Jhaveri K. 2017 Follow-up treatment effects of contingency management and motivational interviewing on substance use: a meta-analysis. *Psychol. Addict. Behav.* **31**, 403–414. (doi:10.1037/adb0000277)
224. DiClemente CC, Corno CM, Graydon MM, Wiprovnick AE, Knoblach DJ. 2017 Motivational interviewing, enhancement, and brief interventions over the last decade: a review of reviews of efficacy and effectiveness. *Psychol. Addict. Behav.* **31**, 862–887. (doi:10.1037/adb0000318)
225. O'Keefe J, Dostrovsky J. 1971 The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely moving rat. *Brain Res.* **34**, 171–175. (doi:10.1016/0006-8993(71)90358-1)
226. Taube JS, Muller RU, Ranck Jr JB. 1990 Head-direction cells recorded from the postsubiculum in freely moving rats. I. Description and quantitative analysis. *J. Neurosci.* **10**, 420–435. (doi:10.1523/JNEUROSCI.10-02-00420.1990)
227. Taube JS. 1995 Head direction cells recorded in the anterior thalamic nuclei of freely moving rats. *J. Neurosci.* **15**, 70–86. (doi:10.1523/JNEUROSCI.15-01-00070.1995)
228. Skaggs WE, Knierim JJ, Kudrimoti HS, McNaughton BL. 1995 A model of the neural basis of the rat's sense of direction. In *Advances in neural information processing systems 7* (eds G Tesauro, DS Touretzky, TK Leen), pp. 173–180. Cambridge, MA: MIT Press.
229. Redish AD, Elga AN, Touretzky DS. 1996 A coupled attractor model of the rodent head direction system. *Netw. Comput. Neural Syst.* **7**, 671–685. (doi:10.1088/0954-898X_7_4_004)
230. Zhang K. 1996 Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: a theory. *J. Neurosci.* **16**, 2112–2126. (doi:10.1523/JNEUROSCI.16-06-02112.1996)
231. Knierim JJ, Kudrimoti HS, McNaughton BL. 1998 Interactions between idiothetic cues and external landmarks in the control of place cells and head direction cells. *J. Neurophysiol.* **80**, 425–446. (doi:10.1152/jn.1998.80.1.425)
232. Gallistel CR. 1990 *The organization of learning*. Cambridge, MA: MIT Press.
233. Margules J, Gallistel CR. 1988 Heading in the rat: determination by environmental shape. *Anim. Learn. Behav.* **16**, 404–410. (doi:10.3758/BF03209379)
234. Cheng K. 1986 A purely geometric module in the rat's spatial representation. *Cognition* **23**, 149–178. (doi:10.1016/0010-0277(86)90041-7)
235. Hermer L, Spelke ES. 1994 A geometric process for spatial reorientation in young children. *Nature* **370**, 57–59. (doi:10.1038/370057a0)
236. Wang RF, Hermer L, Spelke ES. 1999 Mechanisms of reorientation and object localization by children: a comparison with rats. *Behav. Neurosci.* **113**, 475–485. (doi:10.1037/0735-7044.113.3.475)
237. Widge AS *et al*. 2017 Treating refractory mental illness with closed-loop brain stimulation: progress towards a patient-specific transdiagnostic approach. *Exp. Neurol.* **287**, 461–472. (doi:10.1016/j.expneurol.2016.07.021)
238. American Psychiatric Association. 2013 *Diagnostic and statistical manual of mental disorders (DSM-5®)*. Arlington, VA: American Psychiatric Publishing.
239. Regier DA, Narrow WE, Clarke DE, Kraemer HC, Kuramoto SJ, Kuhl EA, Kupfer DJ. 2013 DSM-5 field trials in the United States and Canada, part II: test-retest reliability of selected categorical diagnoses. *Am. J. Psychiatry* **170**, 59–70. (doi:10.1176/appi.ajp.2012.12070999)
240. Cuthbert BN, Insel TR. 2013 Toward the future of psychiatric diagnosis: the seven pillars of RDoC. *BMC Med.* **11**, 126. (doi:10.1186/1741-7015-11-126)
241. Grisanzio KA, Goldstein-Piekarski AN, Wang MY, Rashed Ahmed AP, Samara Z, Williams LM. 2018 Transdiagnostic symptom clusters and associations with brain, behavior, and daily function in mood, anxiety, and trauma disorders. *JAMA Psychiatry* **75**, 201. (doi:10.1001/jamapsychiatry.2017.3951)
242. Gillan CM, Kosinski M, Whelan R, Phelps EA, Daw ND. 2016 Characterizing a psychiatric symptom dimension related to deficits in goal-directed control. *Elife* **5**, e11305. (doi:10.7554/elife.11305)
243. Dang J, King KM, Inzlicht M. 2020 Why are self-report and behavioral measures weakly correlated? *Trends Cognit. Sci.* **24**, 267–269. (doi:10.31234/osf.io/v796c)
244. Enkavi AZ, Eisenberg IW, Bissett P, Mazza GL, MacKinnon D, Marsch L, Poldrack R. 2019 A large-scale analysis of test-retest reliabilities of self-regulation measures. *PsyArXiv*. (doi:10.31234/osf.io/x5pm4)
245. Poldrack RA, Yarkoni T. 2016 From brain maps to cognitive ontologies: informatics and the search for mental structure. *Annu. Rev. Psychol.* **67**, 587–612. (doi:10.1146/annurev-psych-122414-033729)
246. Gillan CM *et al*. 2020 Comparison of the association between goal-directed planning and self-reported compulsivity vs obsessive-compulsive disorder diagnosis. *JAMA Psychiatry* **77**, 77–85. (doi:10.1001/jamapsychiatry.2019.2998)
247. Basu I *et al*. 2020 Closed loop enhancement and neural decoding of human cognitive control. *bioRxiv*, 2020.04.24.059964. (doi:10.1101/2020.04.24.059964)