PROCEEDINGS B

royalsocietypublishing.org/journal/rspb

Review

Cite this article: Hills TT. 2019 Neurocognitive free will. *Proc. R. Soc. B* **286**: 20190510.
http://dx.doi.org/10.1098/rspb.2019.0510

Received: 22 March 2019
Accepted: 11 July 2019

Subject Category:
Neuroscience and cognition

Subject Areas:
cognition, neuroscience

Keywords:
free will, cognitive control, self, consciousness, libertarianism, compatibilism

Author for correspondence:
Thomas T. Hills
e-mail: t.t.hills@warwick.ac.uk

THE ROYAL SOCIETY PUBLISHING

Neurocognitive free will

Thomas T. Hills

University of Warwick, Gibbet Hill Road, Coventry CV4 7AL, UK

TTH, 0000-0003-3842-2076

Free will is an apparent paradox because it requires a historical identity to escape its history in a self-guided fashion. Philosophers have itemized design features necessary for this escape, scaling from action to agency and vice versa. These can be organized into a coherent framework that neurocognitive capacities provide and that form a basis for neurocognitive free will. These capacities include (1) adaptive access to unpredictability, (2) tuning of this unpredictability in the service of hierarchical goal structures, (3) goal-directed deliberation via search over internal cognitive representations, and (4) a role for conscious construction of the self in the generation and choice of alternatives. This frames free will as a process of generative self-construction, by which an iterative search process samples from experience in an adaptively exploratory fashion, allowing the agent to explore itself in the construction of alternative futures. This provides an explanation of how effortful conscious control modulates adaptive access to unpredictability and resolves one of free will's key conceptual problems: how randomness is used in the service of the will. The implications provide a contemporary neurocognitive grounding to compatibilist and libertarian positions on free will, and demonstrate how neurocognitive understanding can contribute to this debate by presenting free will as an interaction between our freedom and our will.

1. Introduction

Free will can be defined as the ability to be free from one's past and yet to simultaneously act in accordance with one's will. This is an apparent paradox because it requires that free will be an ahistorical process governed by a historical identity. The goal of this article is to show that by grounding the design features of free will called for by philosophers in a contemporary understanding of neurocognitive capacities we can resolve this paradox and understand how free will could work in a neurocognitive system.

The design features of free will have been proposed by two primary philosophical camps: the *compatibilists*—who hold that free will is compatible with a deterministic universe—and the *libertarians*—who hold that free will is compatible with an indeterministic universe. The classical compatibilists (such as Hobbes, Locke, Leibniz, Mill and Hume) claim that an agent has free will if it could have done otherwise. The neo-compatibilists (such as Frankfurt, Watson and Bratman) claim that an agent has free will when lower-order desires are aligned with higher-order desires, sometimes called 'wanting what you want', which invokes a hierarchy of wants and volitions that some argue presupposes a wanting 'self'. The libertarians (such as Descartes and Kant) claim that an agent has free will if this conscious self can rationally escape the causal certainties of determinism [1,2].

Discussions of free will have thus far failed to describe how these requirements could be met by a neurocognitive system. This has hampered progress on understanding how free will could work and enfeebled our understanding of ourselves at the same time. As a result, many have provocatively claimed that free will is an illusion because unconscious neural activity at least sometimes precedes or influences conscious activity [3–5]. If consciousness is bypassed in the decision-making process then the worry is 'that the causal chain leading up to our actions bypasses the self' [6], eliminating the kind of free will that many people want [7,8]. Indeed, many of these provocative claims point to

unconscious 'bottom-up' control and use it to argue that a system has no conscious 'top-down' control (for a variety of arguments on both sides see [4,9–13]).

However, without an explanation of how a biological system could satisfy the design features of free will, we risk logical fallacies based on mistaking parts of the system for the whole. To overcome this, we first need to understand how free will could work as a set of neurocognitive processes. We can then assess free will in relation to an operationalized architecture that is grounded in biological reality—treating free will as a quantitative trait [14,15].

Before getting to the design features, it is useful to first define what I mean by conscious control as we will revisit this idea throughout. Conscious control processes are effortful, they focus attention in the face of interference, they experience information in a serial format (one thing at a time), they can generate solutions that are not hard-wired, and they operate over a constrained cognitive workspace—working memory—to which 'we' have access and can later report on as a component of conscious awareness [16–20]. When additional tasks are added to consciously effortful tasks performance suffers. Effortful processes sit in contrast to automatic processes, which are fast and parallel, and do not require conscious awareness. Effortful tasks can be made automatic through repetition (like reading and driving [21]) and when they become automatic they suffer less from the addition of a secondary task. Effortful and automatic processes are typically thought to sit at opposite ends of a continuum and the evidence provided below shows that they can influence one another.

The relationship between effortful processing (sometimes called executive processing) and conscious control is well documented (e.g. [16,18,22]). If we identify effortful consciousness with the self and this effortful self plays a role in satisfying the design features of free will discussed in further detail below, then what people mean by and want from free will are satisfied by our neurocognitive capacities.

2. Adaptive behavioural variability: the principle of alternative possibilities

One of the central tenets of free will held by many philosophers, biologists, legal institutions and common intuition alike is that for an organism to be free it must be the case that it could have done otherwise. This *principle of alternative possibilities* is a core feature of Hobbesian compatibilism [23]. Though some modern views claim that this kind of freedom is irrelevant to free will [24], many still hold that it is necessary and even sufficient [1,14,25].

In order to understand how a cognitive system could generate alternatives, it is first necessary to establish that these alternatives exist. In this section, I will describe the evidence for alternative generation. In the next section, I will discuss where the variability underlying these alternatives originates and how it is controlled. The key point is that organisms can adaptively tune the behavioural variability necessary to generate alternatives and that, at least in humans, this is aided by processes associated with conscious effortful control.

Behavioural variability refers to an organism's ability to produce a variety of alternative responses to the same circumstances, situation or stimuli. There are at least two frequently encountered situations in which this is adaptive: engaging in exploration and outwitting adversaries.

(a) Exploration

Exploration is necessary when solutions to problems are unknown or inadequate and there is not time or the cognitive capacity to systematically examine all the possibilities. In these cases, behavioural variability allows organisms to explore the solution space by trying out different alternatives. In doing so, behavioural variability allows organisms to identify and track favourable locations in the fitness space more quickly than genetic adaptation alone [26].

When organisms forage in resource environments that are clumpy (spatially or temporally autocorrelated), tunable search strategies mediate the trade-off between exploration and exploitation by, for example, adding or subtracting variability from their turning angle. This allows organisms to track resource density even when they do not know exactly where the resources are [27,28]. A relatively low-complexity example is the *run-and-tumble behaviour* of *Escherichia coli*, which moves up resource gradients by randomly choosing a new direction if the current direction is moving the bacterium into a lower quality environment [29]. *Area-restricted search* is a similar behaviour found across metazoans (e.g. vertebrates and invertebrates), with organisms restricting search around locations where they have recently found resources but expanding the range over which random directions are explored when the present environment becomes undesirable [30,31].

This is useful in problem-solving as well. When a solution is unavailable because of ignorance of the causal architecture of the environment, organisms often try effectively random solutions. A classic example is the hungry cats Edward Thorndike placed in locked puzzle boxes [32]. To escape and gain access to food, the cats initially tried (random) sequences of actions. After gaining experience with a box, behaviour became more systematic and the cats escaped more quickly. In a more extreme example, by rewarding porpoises for novel actions they had never before taken, Pryor, Haag & O'Reilly [33] were able to put exploration of new behaviour under stimulus control, making novel behaviour the solution to a consistent stimulus (see also [34]).

Animals often need to maintain exploration even in the face of existing solutions. Sometimes this is because environments change over time [35,36]. Other times it is because finding good solutions requires sufficient exploration of the alternatives [37]. In a standard decision-making paradigm called the Iowa Gambling Task people repeatedly choose cards from four decks and receive differing amounts of money depending on the value of each choice [38]. Two decks are good and two decks are bad. The bad decks do better than the good decks most of the time, leading most people to initially favour them. But they are occasionally associated with large losses, making them worse in the long run. To solve this problem, exploratory behaviour must be maintained even when preliminary evidence suggests the presence of 'better' solutions.

Critical to neurocognitive free will, the modulation of exploration also extends to exploration in internal environments. Retrieving information from memory often requires that one be able to retrieve a variety of answers to the same question. This aids in convergent search—searching for a unique solution such as remembering a person's name

or a word that is common to EVENING, LIGHT and SHOOTING—as well as divergent search—searching for as many solutions as possible, such as items on a shopping list or novel uses for an object [39,40].

In these tasks—which are typical of cognitive search problems—studies suggest that only one idea or memory comes into the workspace of consciousness at a time. As a result, effortful control processes are used to manage the scope over which search operates. This allows individuals to retrieve more or less similar items, producing more or less coherent sequences of recalled information [41–43].

The modulation of internal scope is a kind of area-restricted search in a high-dimensional space that can be characterized as activation patterns in a distributed representation, such as a connectionist or neural network. Such a search can present varying activation patterns as a probe (the question or goal) and then explore patterns of activation in the representation for matches (solutions). The information in the representation is unconscious until it is made conscious by the act of retrieving it and it will be more or less predictable in relation to other knowledge depending on the constraints on exploration.

(b) Outwitting adversaries

Game-theoretic situations between two adversaries often require mixed strategies for which success depends on being unpredictable. Maynard Smith [44] provides the example of two predatory wasps that compete over a hole to lay their eggs. Exposed to a constant threat of injury, each wasp must determine how long to fight over the hole. Maynard Smith demonstrated that the best strategy in this kind of situation is a mixed strategy, with individuals drawing their time in combat from a distribution of values, but not letting their adversary know the outcome of their choice. This is an evolutionarily stable strategy because one individual cannot use information about the other individual's strategy to improve their own. Penalty kick locations in football obey a similar logic of unpredictability [45]. Outwitting adversaries may even require that individuals be capable of being unpredictable even to themselves [46].

Being too predictable is a problem. Prey that behave too predictably are more likely to become the meals of predators who can exploit that predictability. In elegant work describing animal solutions to the Hobbesian capacity to do otherwise, Brembs [14] describes a variety of invertebrates that use randomly initiated evasive maneuvers to escape predation or are alternatively exploited because of predictable behaviour. Cockroaches, for example, select from a repertoire of random directions when escaping predators [47]. Extending this defensive taxonomy to mammals, LeDoux & Daw [48] elaborate on the neural mechanisms underlying defensive behaviour and show its relation to goal-directed behaviour and consciousness, two design features of free will which I discuss in subsequent sections.

In modelling many of the behaviours described above, researchers use probabilistic (softmax) choice rules that incorporate *sensitivity* or *inverse temperature* parameters to modulate variability (e.g. [49,50]). Moreover, this modulation is observed to change over the course of learning such that choices often become less variable with more experience [51].

The above evidence indicates that behavioural variability is adaptive, in the behavioural repertoire of a wide variety of organisms and under internal control, with organisms able to increase or decrease variability in response to context. Understanding where this variability originates and what it gives an organism access to in the service of neurocognitive free will are the topics addressed next.

3. The origin and control of behavioural variability

Whether or not we describe a system as 'random' often depends on whether we see it as arising from deterministic (pseudo-random) or indeterministic sources. Neurocognitive free will offers inroads for each of these sources. There is a finite precision on cognitive abilities, which is a result of a trade-off between computational accuracy and the metabolic cost of information processing (e.g. [52]). This can lead to sensory noise when information from external stimuli is transformed into a neural representation [53,54]. At smaller scales, neural precision is limited by channel noise—the random opening and closing of ion channels—and synaptic noise—derived from probabilistic vesicle release and the random motion of ligand-gated ion channels [55].

The above mentioned compatibilist sources of noise are consistent with a deterministic universe and may be all that cognition has access to when it *turns up the noise*. Nonetheless, a complete discussion of neurocognitive free will cannot yet discount the possibility that neural systems amplify quantum indeterminism [14,56,57]. Neural systems are commonly characterized as having a sensitive dependence on initial conditions of arbitrarily small size [58,59]. If 'arbitrarily small' includes quantum level influences (see [57,60,61]), then two brains wired such that they would forever remain identical in a deterministic universe could eventually diverge in an indeterministic universe.

Perhaps ironically, neurocognitive free will localizes the volumes that have been written comparing compatibilist and libertarian free will to this rather subtle distinction of where the noise comes from. This may not matter for adaptive purposes [62]. Unless my adversary has the omniscience of *Laplace's demon*—who can perfectly predict all deterministic futures—then the ability to amplify quantum noise is not ultimately necessary to outwit adversaries or explore, but it may nonetheless satisfy architectural constraints on building minds like ours.

What matters more for free will is where the decision to modulate variability comes from. If conscious control in any way influences unpredictability, then consciousness is in the loop that governs future behaviour. One way to examine this is to investigate the ability to generate unpredictable behaviour when conscious control is impaired. Consider random number sequence generation tasks, where people are asked to produce unpredictable sequences (e.g. [63]). If individuals in a random sequence generation task are simultaneously exposed to another task that competes for effortful attention—such as *n*-back tasks requiring memory for an ever-changing sequence of letters—their random sequences become increasingly predictable (e.g. [64,65]). People under time pressure or who suffer from unwanted thoughts also produce more predictable sequences [66], as do individuals with impairment in areas of the brain associated with executive control [63,67,68]. This evidence strongly implicates effortful conscious control in the mediation of unpredictability,

whatever its source.[1] The question we now face is how this unpredictability is used in the service of the will.

4. Hierarchical goals and wanting what you want

Neo-compatibilist theories require that for you to have free will you must 'want what you want'. This is often interpreted as having higher-order wants that can select among lower-order wants. Frankfurt [24] elaborated on this idea by postulating a hierarchical theory of free will. In this hierarchy, second-order volitions have preferences over first-order desires. Theories that involve a consistent mapping among a hierarchy of desires are called *mesh theories*. They claim that when desires at different levels in the cognitive system 'mesh in a harmonious way, then the person acts unimpeded; her actions and the desires or intentions causing them are a free outcome of her exercising her own agency' ([70], p. 177).

By invoking higher-order agency, mesh theories emphasize the importance of an enduring historical agent from which desires, values, and preferences originate. For example, Watson's structural theory is a mesh theory that requires a mapping between a low-level motivational system and a high-level value system, so that agents can want what they value [71]. Bratman's [72] planning theory proposes that free agents maintain higher-order volitions over long periods of time.

The need for higher-order agency in mesh theories led Wolf to label them as the *real self-view*. Mesh theories implicitly require the existence of a unified top-level identity, which we notionally think of us as the self [73]. How this top-level identity is formed is central to understanding neurocognitive free will. First, however, we will examine the neurocognitive evidence for hierarchical control.

Cognitively represented goals lie above automatic responses and drive *goal-directed behaviour*. Neuroscientists consider behaviour goal-directed if it involves the neural representation of a goal that is not maintained by external stimuli [74]. For example, in the match-to-sample task, an animal is required to remember the features of a target object, during a period when it is not present, which the animal must later identify. Studies involving the match-to-sample task demonstrate that goal-directed behaviour involves the prefrontal cortex and that damage to this area leads to impaired performance [75].

Hierarchical goal representations allow for one to maintain high-level goals while exploring over subgoals [74,76,77]. An individual can represent the goal of 'making coffee' while activating subgoals such as heating water, locating coffee grounds, finding a cup, and so on. When a subgoal remains unresolved ('Where is the milk?'), goal maintenance allows for an increase in cognitive variability to search over alternative subgoals ('I could go to the store or have it black'). Regions of the prefrontal cortex are activated when individuals maintain a top-level goal while actively processing subgoals [78,79].

Computational simulations of goal-directed cognition routinely invoke hierarchical goal structures [80,81]. Goals and the motivations that drive them are updated based on integrated contextual information, which can arise from external or unconscious processes (such as a growing hunger during a mid-afternoon presentation) or conscious plans (such as the deliberate plan to eat ahead of time). In doing so, the system is put in control of itself [82,83]. As Vancouver [84] has carefully argued, we do not need to invoke a homunculus—a little person in our head—to generate control theory level explanations of behaviour and motivation. There is no need for an agent within the agent. Even though aspects of the system are capable of conscious self-report and can operate on other parts of the system, they can be operated on in return. To use Gazzaniga's [85] example, as enzymes can modulate the genes that encode them, effortful consciousness can modulate the unconscious processes that influence it.

At this point, we have identified neurocognitive faculties that modulate between exploration and exploitation in the pursuit and satisfaction of hierarchical goals. Conscious executive function operates at both ends of the continuum between goal maintenance—exploitation—and alternative generation—exploration. In the next section, we examine how this enables deliberation.

5. Deliberation

According to Wolf [73] 'One wants to be able to choose in light of the knowledge of one's options and in light of the comparative reasons for and against these options' (p. 92). Wolf [73] claimed that these reasons are acquired through rational deliberation. This involves investigating alternatives to satisfying a goal and accessing cognitive information about those alternatives. Evidence for this was provided above in relation to memory retrieval from distributed representations. But the evidence goes much further.

Using electrical recording from hippocampal place cells—which are active when animals are in specific locations—neuroscientists observe that animals selectively replay patterns of neural activation associated with past experiences [86]. One of the earliest of these studies showed that spatio-temporal patterns of activation in the hippocampus that occurred while awake were later replayed during sleep, as apparent 'dreams' [87]. Researchers now routinely observe neural replay like this in awake animals prior to decision-making, what is called *episodic future thinking* (e.g. [88–90]).

When animals experience a choice point in a decision paradigm this replay of neural activation sweeps ahead in a way that corresponds to a spatially congruent series in front of the animal's actual position. These terminate with activity in the ventral striatum—associated with reward valuing—in proportion to the animal's past experiences of rewards along that route [91]. Alternative routes can be explored during this replay period, but only one route is explored at a time [86]. Novel routes can also be explored, creatively constructing experiences from piecemeal information about past experiences [92], corresponding to our common understanding of deliberation.

The scope of this deliberation is tuned to the demands of the context. Episodic future thinking is highest when animals have limited experience with an environment. With additional experience, animals activate fewer forward sweeps and begin to take more favoured routes [91]. Episodic future thinking also activates longer sweeps to reach goals that are further away, suggesting that it is activated in pursuit of specific goals [93].

These deliberations represent different possible future selves [94]. People use the hippocampus and prefrontal

cortex, among other areas, to imagine these future events [95] and this activation directly precedes decision-making [96]. Correspondingly, patients with hippocampal damage, like H.M. [97] and D.B. [98], explicitly report difficulties in imagining future events (see also [99]). Additionally, Haggard [20] has argued that other areas strongly associated with volition, such as the pre-supplementary motor area, are associated with consciousness prior to actions, inhibition of automatic responses, and the flexibility necessary to achieve challenging goals by creatively combining elementary actions into complex constructions. Conscious goal maintenance would, therefore, appear to guide deliberation and allow comparative exploration of alternatives based on their anticipated outcomes.

In sum, neurocognitive processes have the capacity to deliberate over alternative courses of action using goal-directed exploration of alternatives constructed from internal representations.

6. Constructing the self

Scientific research has produced no shortage of entities which we may consider the self. Broadly speaking, these are well classified by James's [100] division of self-knowledge into the experience of 'I' and the awareness of 'me.' The 'I' is composed of pre-reflective experiences associated with cortical midline structures—the *efference copy*—which may allow individuals to coordinate action across the body [101–103]. 'Me' is something of which we are reflectively aware. It is composed of thoughts about ourselves, which include memories and simulated future events. It is sometimes called *access* or *autonoetic consciousness* and, when maintained over time, gives rise to the narrative or remembering self [16,101,104,105]. For the purposes of free will, 'I' is insufficient. When we experience our bodies acting in contrast to our conscious desires, we do not consider ourselves to be acting of 'our' own accord. Rather, it is only when our 'I' does the bidding of 'me' that we consider our actions to originate from our self and to be of our own free will.

Where does the capacity for thinking about 'me' come from? Dawkins ([106], p. 59) suggested the following:

> Survival machines that can simulate the future are one jump ahead of survival machines who can only learn on the basis of overt trial and error … The evolution of the capacity to simulate seems to have culminated in subjective consciousness … Perhaps consciousness arises when the brain's simulation of the world becomes so complete that it must include a model of itself.

Churchland [107] proposed similarly that representing the world involves a representation of one's self. There is growing evidence to support this view.

Consider the following problem. A host of neuroimaging studies show that brains can activate neural structures associated with action while suppressing execution of that action [108–110]. But if this is the case, then adaptive cognition requires distinguishing between real and simulated experiences of the self. Cognitive architectures that fail to do this *reality monitoring* create false memories, delusions, and hallucinations [111–114]. Indeed, we routinely fail to make this distinction during dreams. The ability to do this at all gives rise to the autonoetic experience of 'me.' Hills & Butterfill [115] labelled the process that allows one to distinguish between simulator and simulated a self-actuating model. It is a knowledge of what is and what is not the 'actual' self.

This brings us to a central point in the free will debate: The effortful conscious processes that an organism experiences as the self making the decision are informed by the evidence for alternatives constructed during deliberation. They are *informed* because one does not have to identify oneself with everything that comes to mind. Components of effortful conscious control, executive inhibition, allow us to explicitly *avoid* acting on everything that comes to mind [116]. Effortful conscious processes can demand a stream of novel constructions until something better comes along, even as those constructions inform the processes that summon them.

Though effortful conscious processes are undoubtedly an emergent outcome of multiple underlying contributors, they are not just an interpreter of upward causation [85]. Through inhibition and goal-directed queries, they construct the self and its alternatives. These alternatives are sampled by ourselves from ourselves. They are not random alternatives introduced from outside the system, but effortful random access to alternatives constructed from our own prior experiences.

The material for these constructed deliberations are sampled in proportion to their past frequency of occurrence and how well they match the present situation over which we are deliberating [117]. This self-sampling is the basis of the empirically supported *self-perception theory* [118,119], which posits that we construct attitudes, preferences, and emotions by observing our own past behaviour. Because the self is sampled more or less stochastically in proportion to the size of the net thrown by effortful consciousness, we may never sample the same self twice. The self is constructed in real-time in response to the present circumstances [120].

A key prediction of this framework is that damage to areas of the brain associated with episodic future thinking (i.e. the hippocampus) should also be associated with an impaired sense of self. Indeed, patients with hippocampal damage report having a degraded sense of self [121,122]. For any healthy individual, whatever choice is made following deliberation is supported by evidence from the sampled construction of memory, which we identify as a product of the self.

7. Discussion

At Darwin College in 1977, Karl Popper gave the first Darwin Lecture. In it he said that he had changed his mind about free will: 'The selection of a kind of behaviour out of a randomly offered repertoire may be an act of choice, even an act of free will … the selection may be from some repertoire of random events, without being random in its turn' ([123], p. 348).

Two-stage accounts of free will like Popper's have been variously proposed for compatibilist and libertarian accounts (e.g. [124,125]). The problem has been understanding how agents might use Popper's 'repertoire of random events' without losing themselves in the process. Neurocognitive free will provides a solution: randomness is wielded by the self on the self, both in generating internal representations and in choosing among them. Effortful conscious processes differentially access material from unconscious representations, creatively construct alternatives, value them, and

then choose among these different possible future selves. As a consequence, neurocognitive free will offers a dignity to both the history of the self and the liberty of its alternatives.

By grounding free will in a neurocognitive reality, neurocognitive free will makes numerous contributions to our understanding of free will. First, it provides a neurocognitive basis for the design features of free will set out by prominent philosophical accounts. These include the capacity to do otherwise, the maintenance of hierarchical goal structures that allow an individual to want what they want, and the capacity for deliberation to search over alternative futures and reconstruct a historical identity that selects among those alternatives. In humans at least, these features have properties associated with effortful consciousness, which facilitates the injection of some form of randomness into cognition, the maintenance of goals, and the inhibition of automatic actions necessary to sample alternatives through deliberation. In doing so, neurocognitive free will provides an answer to the paradox set out in the introduction as to how a historical agent can gain degrees of release from its history and do so in a self-governed fashion.

Neurocognitive free will also provide the neurocognitive architecture necessary to answer the call for free will as a quantitative trait grounded in a naturalistic understanding of modern science [9,14]. Indeed, the features described here combined with future capacities for quantum exploration would give robots a variety of libertarian free will (see [126]), for further exploration of this fascinating idea).

This architecture demonstrates that in actualizing free will one must trade-off the will—the coherence of the constructed self-sampled identity—with freedom—the deliberate introduction of randomness into the construction process. At the ends of this trade-off, we have either no freedom or no will. Many arguments for and against free will have focused on exactly these extremes, to little effect. What neurocognitive free will shows us is that to resolve the paradox—to have the 'free' and the 'will'—one must allow the causal constraints of the past and their deliberate subversion to exist at the same time. Possibly in doing so, neurocognition employs Hobesian free will to better enjoy Frankfurtian free will, but much more needs to be said here.

Neurocognitive free will thus highlights the importance of moving beyond free will as a binary trait that does or does not exist. Attempting to shoehorn a quantitative trait into a categorical definition leads to paradoxical conclusions that violate the intuitions of people on all sides of the issue [7,8,11]. Neurocognitive free will addresses this by reframing free will as a group of individually adaptive traits, which collectively represent a process of *generative self-construction*—goal-directed exploration of alternative future selves that are constructed by the agent from the agent in a creative process akin to the generative construction of language.

Variations of the traits underlying neurocognitive free will satisfy different definitions of free will to different degrees. These variations invite us to ask how different individuals may have more or less free will based on differential capacities for a coherent will in the form of a self-identity, goal maintenance in the face of distractions, a rich cognitive representation capable of offering the materials necessary to compose a wide degree of alternatives, and potentially even different degrees of access to indeterminism.

When access to indeterminism exists—now or in the future [126]—neurocognitive free will allows organisms to explore alternatives that are not causally necessitated by history and choose among them from the view of a rational self. No antecedent event requires just that action, nor is that action chosen outside the sovereignty of a conscious rational self.

By satisfying these two capacities neurocognitive free will provides a mechanism for one of the most challenging requirements of free will, what libertarians call self-forming actions. *Self-forming actions* are required for an organism to escape its own history and do so in a self-guided fashion in an indeterministic world where alternative actions are genuinely possible without any historical differences [7,125]. The difficulty with this requirement has been that prior work could not provide a satisfactory answer as to how the indeterminism necessary to generate these alternatives could be put to use by the self (e.g. [7,57]). Neurocognitive free will shows that deliberation uses conscious cognitive control to modulate the degree of random—and possibly indeterministic—sampling from unconscious representations. In other words, controlled access to a random process is what makes the alternative versions of our future selves possible.

Neurocognitive free will also provides the capacity to rationally support different possible alternative futures. As Kane [1] suggests, what we want from free will is the power to go more than one way 'rationally and deliberately, rather than flukishly, given exactly the same prior deliberation and thought processes' (p. 108). In neurocognitive free will the 'deliberation and thought processes' are where the *more-than-one-way rationality* happens—our rationale is generated in tandem with the alternatives constructed from our cognitive representations.

Even if what comes to mind is influenced by indeterminism, it does not go without consideration by more willful processes [127]. An individual does not get lucky by making a decision consistent with their identity as some critics of libertarianism have claimed [11,128]. Healthy individuals always make the choice most consistent with their identity because they construct that identity in the process of deliberation.

To close, let me point out that the feeling of freedom necessarily comes in part from not knowing ahead of time what we are going to do. Velleman [129] calls this epistemic freedom. It is this lack of knowing, caused by discovering different versions of oneself as one goes along, that has led many to suggest there was no 'real' self in the system to begin with—and therefore no free will. The present accounting turns this thinking on its head. It does so by suggesting it is exactly the finding out—the initiation of the search and the choice among alternatives—that is the basis of the self's emergent will and its genuine freedom. The bringing forth of a self-identity is the evaluation of alternatives through self-simulation. If a historical self emerges through conscious deliberation, and that deliberation involves simulation of alternative futures over which the self chooses, then a historical identity and the capacity for free choice arise in tandem.

Data accessibility. This article has no additional data.

Competing interests. I declare I have no competing interests.

Funding. This work was supported by the Royal Society Wolfson Research Merit Award (WM160074) and the Alan Turing Institute.

Acknowledgements. Generous feedback on earlier drafts of this work were provided by numerous colleagues, students and anonymous reviewers. I would like to particularly thank Sara Hills, Martin Rieke, Steve Butterfill, John Michael, Kita Sotaro, John Pickering, Peter

Todd, Kenneth Schaffner, Nick Watkins, Friederike Schlaghecken, Cynthia Siew, Eva Jimenez Mesa, Li Ying and Tomas Engelthaler.

Endnote

[1]This offers an entry point for free will into the Libet *et al*. [69] studies. A common interpretation of these studies is that because people only become consciously aware of a decision after a random unconscious process 'decides,' consciousness is not involved in decision making (e.g. [11]). This only recognizes one direction of the two-way interaction between conscious and unconscious processes. Neurocognitive free will predicts that consciousness initiates the unconscious random process in the first place in roughly the same way it would initiate an exploratory search for names for a new puppy.

References

1. Kane R. 1998 *The significance of free will*. Oxford, UK: Oxford University Press.
2. Searle JR. 2013 *Freedom and neurobiology: reflections on free will, language, and political power*. New York: Columbia University Press.
3. Crick F, Clark J. 1994 The astonishing hypothesis. *J. Conscious. Stud.* **1**, 10–16.
4. Wegner DM. 2002 *The illusion of conscious will*. Cambridge, MA: MIT Press.
5. Burns K, Bechara A. 2007 Decision making and free will: a neuroscience perspective. *Behav. Sci. Law* **25**, 263–280. (doi:10.1002/bsl.751)
6. Knobe J, Nichols S. 2011 Free will and the bounds of the self. In *The Oxford handbook of free will* (ed. R Kane), pp. 530–554. Oxford, UK: Oxford University Press.
7. Dennett DC. 2003 *Freedom evolves*. London, UK: Viking.
8. Nahmias E. 2015 Why we have free will. *Sci. Am.* **312**, 77–79. (doi:10.1038/scientificamerican0315-77)
9. Baumeister RF. 2008 Free will in scientific psychology. *Perspect. Psychol. Sci.* **3**, 14–19. (doi:10.1111/j.1745-6916.2008.00057.x)
10. Stillman TF, Baumeister RF, Mele AR. 2011 Free will in everyday life: autobiographical accounts of free and unfree actions. *Phil. Psychol.* **24**, 381–394. (doi:10.1080/09515089.2011.556607)
11. Caruso GD. 2012 *Free will and consciousness: a determinist account of the illusion of free will*. Lanham, MD: Lexington Books.
12. Roskies A. 2012 How does the neuroscience of decision making bear on our understanding of moral responsibility and free will? *Curr. Opin Neurobiol.* **22**, 1022–1026. (doi:10.1016/j.conb.2012.05.009)
13. Murphy N, Ellis GFR, O'Connor T. 2009 *Downward causation and the neurobiology of free will*. Berlin, Germany: Springer.
14. Brembs B. 2011 Towards a scientific concept of free will as a biological trait: spontaneous actions and decision-making in invertebrates. *Proc. R. Soc. B* **278**, 930–939. (doi:10.1098/rspb.2010.2325)
15. Lavazza A. 2016 Free will and neuroscience: from explaining freedom away to new ways of operationalizing and measuring it. *Front. Hum. Neurosci.* **10**, 262. (doi:10.3389/fnhum.2016.00262)
16. Wheeler MA, Stuss DT, Tulving E. 1997 Toward a theory of episodic memory: the frontal lobes and autonoetic consciousness. *Psychol. Bull.* **121**, 331–354. (doi:10.1037/0033-2909.121.3.331)
17. Dehaene S, Naccache L. 2001 Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition* **79**, 1–37. (doi:10.1016/S0010-0277(00)00123-2)
18. Baddeley A. 2007 *Working memory, thought, and action*. Oxford, UK: Oxford University Press.
19. Ardila A. 2016 Is 'self-consciousness' equivalent to 'executive function'? *Psychol. Neurosci.* **9**, 215. (doi:10.1037/pne0000052)
20. Haggard P. 2008 Human volition: towards a neuroscience of will. *Nat. Rev. Neurosci.* **9**, 934–946.
21. Schneider W, Shiffrin RM. 1977 Controlled and automatic human information processing: I. Detection, search, and attention. *Psychol. Rev.* **84**, 1–66. (doi:10.1037/0033-295X.84.1.1)
22. Christensen W, Sutton J, McIlwain DJ. 2016 Cognition in skilled action: meshed control and the varieties of skill experience. *Mind Lang.* **31**, 37–66. (doi:10.1111/mila.12094)
23. Hobbes T. 1839–1845 *The English works of Thomas Hobbes of Malmesbury* (ed. SW Molesworth). London, UK: J. Bohn.
24. Frankfurt HG. 1971 Freedom of the will and the concept of a person. *J. Phil.* **68**, 5–20. (doi:10.2307/2024717)
25. Heisenberg M. 2009 Is free will an illusion? *Nature* **459**, 164–165. (doi:10.1038/459164a)
26. Hinton GE, Nowlan SJ. 1987 How learning can guide evolution. *Complex Syst.* **1**, 495–502.
27. Plank M, James A. 2008 Optimal foraging: Lévy pattern or process? *J. R. Soc. Interface* **5**, 1077–1086. (doi:10.1098/rsif.2008.0006)
28. Viswanathan G, Luz M, Raposo E. 2011 *The physics of foraging: an introduction to random searches and biological encounters*. Cambridge, UK: Cambridge University Press.
29. Macnab RM, Koshland Jr DE. 1972 The gradient-sensing mechanism in bacterial chemotaxis. *Proc. Natl Acad. Sci. USA* **69**, 2509–2512. (doi:10.1073/pnas.69.9.2509)
30. Bell WJ. 1991 *Searching behavior: the behavioral ecology of finding resources*. London, UK: Chapman & Hall.
31. Hills TT. 2006 Animal foraging and the evolution of goal-directed cognition. *Cogn. Sci.* **30**, 3–41. (doi:10.1207/s15516709cog0000_50)
32. Thorndike EL. 1898 Animal intelligence: an experimental study of the associative processes in animals. *Psychol. Rev. Monogr. Suppl.* **2**, 109.
33. Pryor KW, Haag R, O'Reilly J. 1969 The creative porpoise: training for novel behavior. *J. Exp. Anal. Behav.* **12**, 653–661. (doi:10.1901/jeab.1969.12-653)
34. Neuringer A, Jensen G. 2010 Operant variability and voluntary action. *Psychol. Rev.* **117**, 972–993. (doi:10.1037/a0019499)
35. Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ. 2006 Cortical substrates for exploratory decisions in humans. *Nature* **441**, 876–879. (doi:10.1038/nature04766)
36. Stephens DW, Krebs JR. 1987 *Foraging theory*. Princeton, NJ: Princeton Academic Press.
37. Worthy DA, Gorlick MA, Pacheco JL, Schnyer DM, Maddox WT. 2011 With age comes wisdom: decision making in younger and older adults. *Psychol. Sci.* **22**, 1375–1380. (doi:10.1177/0956797611420301)
38. Bechara A, Damasio AR, Damasio H. 1994 Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition* **50**, 7–15. (doi:10.1016/0010-0277(94)90018-3)
39. Smith KA, Huber DE, Vul E. 2013 Multiply-constrained semantic search in the remote associates test. *Cognition* **128**, 64–75. (doi:10.1016/j.cognition.2013.03.001)
40. Hills TT, Todd PM, Lazer D, Redish AD, Couzin ID, the Cognitive Search Research Group. 2015 Exploration versus exploitation in space, mind, and society. *Trends Cogn. Sci.* **19**, 46–54. (doi:10.1016/j.tics.2014.10.004)
41. Rosen V, Engle RW. 1997 The role of working memory capacity in retrieval. *J. Exp. Psychol. Gen.* **126**, 211–227. (doi:10.1037/0096-3445.126.3.211)
42. Unsworth N, Engle RW. 2007 The nature of individual differences in working memory capacity: active maintenance in primary memory and controlled search from secondary memory. *Psychol. Rev.* **114**, 104–132. (doi:10.1037/0033-295X.114.1.104)
43. Hills TT, Mata R, Wilke A, Samanez-Larkin G. 2013 Mechanisms of age-related decline in memory search across the adult life span. *Dev. Psychol.* **49**, 2396–2404. (doi:10.1037/a0032272)
44. Maynard Smith J. 1982 *Evolution and the theory of games*. Cambridge, UK: Cambridge University Press.
45. Chiappori P-A, Levitt S, Groseclose T. 2002 Testing mixed-strategy equilibria when players are heterogeneous: the case of penalty kicks in soccer. *Am. Econ. Rev.* **92**, 1138–1151. (doi:10.1257/00028280260344678)
46. Trivers R. 2011 *Deceit and self-deception: fooling yourself the better to fool others*. London, UK: Penguin.
47. Domenici P, Booth D, Blagburn JM, Bacon JP. 2008 Cockroaches keep predators guessing by using preferred escape trajectories. *Curr. Biol.* **18**, 1792–1796. (doi:10.1016/j.cub.2008.09.062)
48. LeDoux J, Daw ND. 2018 Surviving threats: neural circuit and computational implications of a new

taxonomy of defensive behaviour. *Nat. Rev. Neurosci.* **19**, 269–282. (doi:10.1038/nrn.2018.22)

49. Collins A, Koechlin E. 2012 Reasoning, learning, and creativity: frontal lobe function and human decision-making. *PLoS Biol.* **10**, e1001293. (doi:10.1371/journal.pbio.1001293)
50. Sutton RS, Barto AG. 1998 *Reinforcement learning*. Cambridge, MA: MIT Press.
51. Yechiam E, Busemeyer J, Stout J. 2005 Using cognitive models to map relations between neuropsychological disorders and human decision-making deficits. *Psychol. Sci.* **16**, 973–978. (doi:10.1111/j.1467-9280.2005.01646.x)
52. Drugowitsch J, Moreno-Bote R, Churchland AK, Shadlen MN, Pouget A. 2012 The cost of accumulating evidence in perceptual decision making. *J. Neurosci.* **32**, 3612–3628. (doi:10.1523/JNEUROSCI.4010-11.2012)
53. 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)
54. Kaufman MT, Churchland AK. 2013 Cognitive neuroscience: sensory noise drives bad decisions. *Nature* **496**, 172–173. (doi:10.1038/496172a)
55. White JA, Rubinstein JT, Kay AR. 2000 Channel noise in neurons. *Trends Neurosci.* **23**, 131–137. (doi:10.1016/S0166-2236(99)01521-0)
56. Hobbs J. 1991 Chaos and indeterminism. *Can. J. Phil.* **21**, 141–164. (doi:10.1080/00455091.1991.10717241)
57. Koch C. 2009 Free will, physics, biology, and the brain. In *Downward causation and the neurobiology of free will* (eds N Murphy, GFR Ellis, T O'Connor), pp. 31–52. Berlin, Germany: Springer.
58. Garson JW. 1995 Chaos and free will. *Phil. Psychol.* **8**, 365–374. (doi:10.1080/09515089508573165)
59. Kitzbichler MG, Smith ML, Christensen SR, Bullmore E. 2009 Broadband criticality of human brain network synchronization. *PLoS Comput. Biol.* **5**, e1000314. (doi:10.1371/journal.pcbi.1000314)
60. Chua L. 2014 Memristor, Hodgkin-Huxley, and edge of chaos. *Nanotechnology* **24**, 383001. (doi:10.1088/0957-4484/24/38/383001)
61. Chua L, Sbitnev V, Kim H. 2012 Neurons are poised near the edge of chaos. *Int. J. Bifurcation Chaos* **22**, 1250098. (doi:10.1142/S0218127412500988)
62. Taylor C, Dennett DC. 2011 Who's afraid of determinism? Rethinking causes and possibilities. In *The Oxford handbook of free will* (ed. R Kane), pp. 257–277. Oxford, UK: Oxford University Press.
63. Spatt J, Goldenberg G. 1993 Components of random generation by normal subjects and patients with dysexecutive syndrome. *Brain Cogn.* **23**, 231–242. (doi:10.1006/brcg.1993.1057)
64. Baddeley A, Emslie H, Kolodny J, Duncan J. 1998 Random generation and the executive control of working memory. *Q. J. Exp. Psychol.* **51**, 819–852. (doi:10.1080/713755788)
65. Brugger P, Monsch A, Salmon D, Butters N. 1996 Random number generation in dementia of the Alzheimer type: a test of frontal executive functions. *Neuropsychologia* **34**, 97–103. (doi:10.1016/0028-3932(95)00066-6)
66. Hirsch CR, Mathews A. 2012 A cognitive model of pathological worry. *Behav. Res. Ther.* **50**, 636–646. (doi:10.1016/j.brat.2012.06.007)
67. Artiges E, Salamé P, Recasens C, Poline J-B, Attar-Levy D, De la Raillère A, Paillère-Martinot ML, Danion J-M, Martinot J-L. 2000 Working memory control in patients with schizophrenia: a pet study during a random number generation task. *Am. J. Psychiatry* **157**, 1517–1519. (doi:10.1176/appi.ajp.157.9.1517)
68. Wiegersma S, van der Scheer E, Hijman R. 1990 Subjective ordering, short-term memory, and the frontal lobes. *Neuropsychologia* **28**, 95–98. (doi:10.1016/0028-3932(90)90089-7)
69. Libet B, Gleason CA, Wright EW, Pearl DK. 1983 Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential). *Brain* **106**, 623–642. (doi:10.1093/brain/106.3.623)
70. McKenna M. 2011 Contemporary compatibilism: Mesh theories and reasons-responsive theories. In *The Oxford handbook of free will* (ed. R Kane), pp. 175–196. Oxford, UK: Oxford University Press.
71. Watson G. 1975 Free agency. *J. Phil.* **72**, 205–220. (doi:10.2307/2024703)
72. Bratman M. 1987 *Intention, plans, and practical reason*. Cambridge, MA: Harvard University Press.
73. Wolf S. 1993 *Freedom within reason*. Oxford, UK: Oxford University Press.
74. Winstanley CA *et al.* 2012 Search, goals, and the brain. In *Cognitive search: evolution, algorithms, and the brain* (eds PM Todd, TT Hills, T Robbins), pp. 125–156. Cambridge, MA: MIT Press.
75. Goldman-Rakic PS. 1995 Cellular basis of working memory. *Neuron* **14**, 477–485. (doi:10.1016/0896-6273(95)90304-6)
76. Austin JT, Vancouver JB. 1996 Goal constructs in psychology: structure, process, and content. *Psychol. Bull.* **120**, 338–375. (doi:10.1037/0033-2909.120.3.338)
77. Botvinick M. 2008 Hierarchical models of behavior and prefrontal function. *Trends Cogn. Sci.* **12**, 201–208. (doi:10.1016/j.tics.2008.02.009)
78. Koechlin E, Basso G, Pietrini P, Panzer S, Grafman J. 1999 The role of the anterior prefrontal cortex in human cognition. *Nature* **399**, 148–151. (doi:10.1038/20178)
79. Miller EK, Cohen JD. 2001 An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci.* **24**, 167–202. (doi:10.1146/annurev.neuro.24.1.167)
80. Newell A, Simon HA. 1972 *Human problem solving*. Oxford, UK: Prentice-Hall.
81. Anderson JR, Lebiere CJ. 1998 *The atomic components of thought*. Mahwah, NJ: Erlbaum.
82. Hazy TE, Frank MJ, O'Reilly RC. 2006 Banishing the homunculus: making working memory work. *Neuroscience* **139**, 105–118. (doi:10.1016/j.neuroscience.2005.04.067)
83. Hills TT, Todd PM, Goldstone RL. 2010 The central executive as a search process: priming exploration and exploitation across domains. *J. Exp. Psychol. Gen.* **139**, 590–609. (doi:10.1037/a0020666)
84. Vancouver JB. 2005 The depth of history and explanation as benefit and bane for psychological control theories. *J. Appl. Psychol.* **90**, 38–52. (doi:10.1037/0021-9010.90.1.38)
85. Gazzaniga MS. 2012 *Who's in charge? Free will and the science of the brain*. New York: Ecco.
86. Pezzulo G, van der Meer MA, Lansink CS, Pennartz CM. 2014 Internally generated sequences in learning and executing goal-directed behavior. *Trends Cogn. Sci.* **18**, 647–657. (doi:10.1016/j.tics.2014.06.011)
87. Skaggs WE, McNaughton BL. 1996 Replay of neuronal firing sequences in rat hippocampus during sleep following spatial experience. *Science* **271**, 1870–1873. (doi:10.1126/science.271.5257.1870)
88. Foster DJ, Wilson MA. 2006 Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature* **440**, 680–683. (doi:10.1038/nature04587)
89. Diba K, Buzsáki G. 2007 Forward and reverse hippocampal place-cell sequences during ripples. *Nat. Neurosci.* **10**, 1241–1242. (doi:10.1038/nn1961)
90. Carr MF, Jadhav SP, Frank LM. 2011 Hippocampal replay in the awake state: a potential substrate for memory consolidation and retrieval. *Nat. Neurosci.* **14**, 147–153. (doi:10.1038/nn.2732)
91. Johnson A, van der Meer MA, Redish AD. 2007 Integrating hippocampus and striatum in decision-making. *Curr. Opin Neurobiol.* **17**, 692–697. (doi:10.1016/j.conb.2008.01.003)
92. Pfeiffer BE, Foster DJ. 2013 Hippocampal place-cell sequences depict future paths to remembered goals. *Nature* **497**, 74–79. (doi:10.1038/nature12112)
93. Wikenheiser AM, Redish AD. 2015 Hippocampal theta sequences reflect current goals. *Nat. Neurosci.* **18**, 289–294. (doi:10.1038/nn.3909)
94. Markus H, Nurius P. 1986 Possible selves. *Am. Psychol.* **41**, 954–969. (doi:10.1037/0003-066X.41.9.954)
95. Buckner RL. 2010 The role of the hippocampus in prediction and imagination. *Annu. Rev. Psychol.* **61**, 27–48. (doi:10.1146/annurev.psych.60.110707.163508)
96. Peters J, Büchel C. 2010 Episodic future thinking reduces reward delay discounting through an enhancement of prefrontal-mediotemporal interactions. *Neuron* **66**, 138–148. (doi:10.1016/j.neuron.2010.03.026)
97. Scoville W. 1968 Amnesia after bilateral mesial temporal-lobe excision: introduction to case H.M. *Neuropsychologia* **6**, 211–213. (doi:10.1016/0028-3932(68)90020-1)
98. Klein SB, Loftus J, Kihlstrom JF. 2002 Memory and temporal experience: the effects of episodic memory loss on an amnesic patient's ability to remember the past and imagine the future. *Soc. Cognition* **20**, 353–379. (doi:10.1521/soco.20.5.353.21125)
99. Tulving E. 1985 Memory and consciousness. *Can. Psycho./Psychologie Can.* **26**, 1–12. (doi:10.1037/h0080017)
100. James W. 1890 *The principles of psychology*. New York, NY: Holt.

101. Block N. 1995 How many concepts of consciousness? *Behav. Brain Sci.* **18**, 272–287. (doi:10.1017/S0140525X00038486)
102. Panksepp J. 1998 The periconscious substrates of consciousness: affective states and the evolutionary origins of the self. *J. Conscious. Stud.* **5**, 566–582.
103. Metzinger T. 2004 *Being no one: the self-model theory of subjectivity*. Cambridge, MA: MIT Press.
104. Neisser U, Fivush R. 1994 *The remembering self: construction and accuracy in the self-narrative*. Cambridge, UK: Cambridge University Press.
105. Gallagher S. 2000 Philosophical conceptions of the self: Implications for cognitive science. *Trends Cogn. Sci.* **4**, 14–21. (doi:10.1016/S1364-6613(99)01417-5)
106. Dawkins R. 2016 *The selfish gene*. Oxford, UK: Oxford University Press.
107. Churchland PS. 2002 *Brain-wise: studies in neurophilosophy*. Cambridge, MA: MIT Press.
108. Barsalou LW. 2008 Grounded cognition. *Annu. Rev. Psychol.* **59**, 617–645. (doi:10.1146/annurev.psych.59.103006.093639)
109. Hesslow G. 2012 The current status of the simulation theory of cognition. *Brain Res.* **1428**, 71–79. (doi:10.1016/j.brainres.2011.06.026)
110. Meister IG, Krings T, Foltys H, Boroojerdi B, Müller M, Töpper R, Thron A. 2004 Playing piano in the mind—an fMRI study on music imagery and performance in pianists. *Cogn. Brain Res.* **19**, 219–228. (doi:10.1016/j.cogbrainres.2003.12.005)
111. Johnson MK, Raye CL. 1981 Reality monitoring. *Psychol. Rev.* **88**, 67–85. (doi:10.1037/0033-295X.88.1.67)
112. Morrison AP. 2001 The interpretation of intrusions in psychosis: an integrative cognitive approach to hallucinations and delusions. *Behav. Cogn. Psychother.* **29**, 257–276. (doi:10.1017/S1352465801003010)
113. Brainerd CJ, Reyna VF. 2005 *The science of false memory*. Oxford, UK: Oxford University Press.
114. Mitchell KJ, Johnson MK. 2009 Source monitoring 15 years later: what have we learned from fMRI about the neural mechanisms of source memory? *Psychol. Bull.* **135**, 638–677. (doi:10.1037/a0015849)
115. Hills TT, Butterfill S. 2015 From foraging to autonoetic consciousness: the primal self as a consequence of embodied prospective foraging. *Cur. Zool.* **61**, 368–381. (doi:10.1093/czoolo/61.2.368)
116. Miyake A, Friedman NP, Emerson MJ, Witzki AH, Howerter A, Wager TD. 2000 The unity and diversity of executive functions and their contributions to complex 'frontal lobe' tasks: a latent variable analysis. *Cogn. Psychol.* **41**, 49–100.
117. Sanborn AN, Chater N. 2016 Bayesian brains without probabilities. *Trends Cogn. Sci.* **20**, 883–893. (doi:10.1016/j.tics.2016.10.003)
118. Bem DJ. 1972 Self-perception theory. *Adv. Exp. Soc. Psychol.* **6**, 1–62.
119. Johansson P, Hall L, Tärning B, Sikström S, Chater N. 2014 Choice blindness and preference change: you will like this paper better if you (believe you) chose to read it! *J. Behav. Decis. Making* **27**, 281–289. (doi:10.1002/bdm.1807)
120. Chater N. 2018 *The mind is flat: the illusion of mental depth and the improvised mind*. UK: Penguin.
121. Corkin S. 2002 What's new with the amnesic patient H.M.? *Nat. Rev. Neurosci.* **3**, 153–160. (doi:10.1038/nrn726)
122. Hassabis D, Kumaran D, Vann SD, Maguire EA. 2007 Patients with hippocampal amnesia cannot imagine new experiences. *Proc. Natl Acad. Sci. USA* **104**, 1726–1731. (doi:10.1073/pnas.0610561104)
123. Popper K. 1978 Natural selection and the emergence of mind. *Dialectica* **32**, 339–355. (doi:10.1111/j.1746-8361.1978.tb01321.x)
124. Dennett DC. 1984 *Elbow room: the varieties of free will worth wanting*. Cambridge, MA: MIT Press.
125. Kane R. 1985 *Free will and values*. Albany, NY: SUNY Press.
126. Briegel HJ. 2012 On creative machines and the physical origins of freedom. *Sci. Rep.* **2**, 522. (doi:10.1038/srep00522)
127. Mele AR. 1996 Soft libertarianism and Frankfurt-style scenarios. *Phil. Top.* **24**, 123–141. (doi:10.5840/philtopics199624220)
128. Mele AR. 2008 *Free will and luck*. Oxford, UK: Oxford University Press.
129. Velleman JD. 1989 *Practical reflection*. Princeton, NJ: Princeton Univerity Press.