Accepted for publication at Philosophical Transactions of the Royal Society B
DOI: 10.1098/rstb.2018.0139

Why hasn’t evolution selected for perfect self-control?

Benjamin Y. Hayden

Department of Neuroscience and
Center for Magnetic Resonance Research
University of Minnesota,
Minneapolis, MN 55455

Correspondence:
Benjamin Y. Hayden
CMRR Building,
University of Minnesota,
Minneapolis, MN 55455
Telephone: 425-749-2341
Email: benhayden@gmail.com
ABSTRACT

Self-control refers to the ability to deliberately reject tempting options and instead select ones that produce greater long-term benefits. Although some apparent failures of self-control are, on closer inspection, reward maximizing, at least some self-control failures are clearly disadvantageous and non-strategic. The existence of poor self-control presents an important evolutionary puzzle because there is no obvious reason why good self-control should be more costly than poor self-control. After all, a rock is infinitely patient. I propose that self-control failures result from cases in which well-learned (and thus routinized) decision making strategies yield suboptimal choices. These mappings persist in the decision-makers’ repertoire because they result from learning processes that are adaptive in the broader context, either on the timescale of learning or of evolution. Self-control, then, is a form of cognitive control and the subjective feeling of effort likely reflects the true costs of cognitive control. Poor self-control, in this view, is ultimately a result of bounded optimality.
MAIN TEXT

Introduction

Poor self-control is inimical to mental and physical health and to life success; it is associated with poverty, obesity, loneliness, and other unwanted states (de Ridder et al., 2012; Tangney et al., 2004; Story et al., 2014). It is both a symptom and a cause of diseases that increase mortality, such as addiction, depression, and obsessive-compulsive disorder (e.g. Duckworth and Seligman 2005; Quinn and Fromme 2010; Dixon et al., 2003). Because failures of self-control are costly, the ability to exert self-control can confer evolutionary benefits and ought to be subject to strongly negative selection pressure. The ubiquity of poor self-control, then, poses an important riddle: why hasn’t natural selection endowed us with perfect self-control?

For present purposes, I define self-control as deliberately avoiding the choice of a tempting option so as to choose an alternative that produces greater long-term benefits. The main reason I use this definition is because it brings to the fore the evolutionarily puzzling aspects of self-control failure. This definition is not universally shared, but it is, from a reading of the literature, the closest to a consensus view available (e.g. Shenhav, 2017; Shenhav et al., 2017; Muraven and Baumeister, 2000; Baumeister et al., 1994; Mischel et al., 1996). Some other definitions include strategies that avoiding tempting contexts; my definition treats these as outside the bounds of self-controlled behavior (Fujita, 2011; Berkman et al., 2017). Other scholars have considered choices that appear to reflect poor self-control may have adaptive outcomes (Fawcett et al., 2012; Stephens, 2002; Stevens, 2014; Stephens and Anderson, 2001; Kacelnik, 2003). Such choices are interesting, as they provide insight into the evolution of cognitive faculties. However, they are not relevant to the central question I consider here, as they are reward maximizing, and thus adaptive, and not evolutionarily puzzling. One important caveat in this definition is that to be called a test of self-control, the decision-maker must know the potential outcomes (or the range of outcomes in the case of stochastic decisions).

It’s not obvious why perfect self-control would be difficult to evolve. Self-control decisions are, ostensibly, just like another mental operation. Consider, for comparison,
the example of saccadic eye movements, which, like self-control decisions, are regulated by the brain (including the prefrontal cortex) and are subject to volitional control (Yarbus, 1967). We make saccadic eye movements three to four times per second during our waking hours, without any sense of fatigue, throughout our lives. When something surprising appears in the visual field, we look at it without succumbing to the temptation to procrastinate for several days (as we might with more conventional self-control problems, such as paying a bill or reviewing a manuscript). We don’t ever feel the temptation to cheat or cut corners; for example, we don’t move our eyes only 80% of the way towards a target, as we might with a diet. Our oculomotor control systems, like our respiratory control systems, our form vision systems, and many others, constantly function at a high level with rare failures. So what makes economic decisions different? Why are so many of our daily value-based decisions subject to large and small self-control failures?

**Some approaches to self-control do not help us understand why it fails**

One view of self-control, dating back at least to William James, sees it as the result of the competition between two systems (James, 1890; Baumeister and Heatherton 1996; Carver et al., 2005; Hoffman et al., 2009). These are often known as the hot and cold systems. The hot system advocates for impulsive choices and the cold system advocates for controlled ones (Metcalfe and Mischel, 1999). These two systems have at least some affiliation with Freud’s idea of the id and superego, respectively (Baumeister et al., 1998). And in a more mathematical guise, this view has direct parallels to the beta and delta systems (Harris and Laibson, 2004; McClure et al., 2004). The hot vs. cold idea is supported by neuroscientific results showing a regulatory system (often dorsal and lateral) inhibiting a basic value system (often ventral and medial, e.g. Figner et al, 2010; Hare et al, 2009; McClure et al., 2004). In this two-systems view, self-control failure reflects a failure of the cold system to overrule the hot system. Despite its appeals, this view does not provide any explanation for why the hot system would ever win. That is, it simply allow for a restatement of the core mystery of self-control.
Another important view sees self-control as an economic decision – a comparison between two differently valued options - that is not different in any substantial way from other economic decisions (Berkman et al., 2017; see also McGuire and Kable, 2015; Kurzban et al., 2013; Becker and Murphy, 1988). While self-control decisions clearly are a type of economic decision, they are of a special type. The strict similarity view ignores the most important thing about self-control: it can fail. And those failures are not just due to noise. Many cognitive processes (including economic choice) are susceptible to errors but these errors are due to noise and are independent of choices. In contrast, self-control failures always go in the same direction: succumbing to temptation. This distinction is clearest in the case of intertemporal choice tasks. In standard implementations of these tasks, preference for the shorter-sooner option is most often indicates poor self-control (Kirby and Marakovic, 1996). But on trials in which the shorter-sooner option provides a higher long-term reward rate than the larger-later alternative, then choice of the larger-later option will yield negative discount rates. This ‘negative self-control’ is much rarer than the alternative.

Another limitation of the economic model is that it does not readily explain the ego-dystonic nature of self-control failures. That is, it cannot explain why failure, or even the prospect of it, would evoke negative emotions (often severe ones), when economic mistakes do not (for a similar argument, see Rogeberg, 2004).

Most critically for my concerns here, the economic view cannot explain the high prevalence of self-control failure if it is just a type of miscalculation. Evolution has finely honed our minds to make good decisions (Santos and Rosati, 2015; Stephens and Krebs, 1986). It has endowed us with abilities to do cost-benefit computations that are much more complicated than many self-control problems require. For example, our brains can simultaneously track multiple fluctuating variables at many timescales (Farashahi et al., 2018; Daw et al., 2006; Pearson et al., 2010); we can detect subtle changes in probability (Nassar et al., 2010); we can anticipate others’ strategies several levels deep (Camerer and Ho, 1999). Time biases don’t seem to be the problem either. In foraging tasks, at least, many animals, including humans, can optimize reward rate to within a few
percentage points of optimal (Blanchard et al., 2014; Schweighofer et al., 2006; Stephens and Krebs, 1986).

A third, not entirely distinct, view equates self-control directly with patience or with withholding a response (Stevens and Stephens, 2008; Kacelnik, 2003; Mischel and Ebbeson 2001; Rachlin, 2000). Patience is a time-centric view of self-control, and it equates poor self-control with unwillingness to wait extended periods of time to obtain better rewards. The patience perspective generally equates poor self-control with action and good self-control with inaction. Asking why patience fails involves asking how action (which would presumably be costly) accidently overcomes inaction. Prima facie, not moving one’s muscles would seem to require very minimal amounts of energy. (This ignores the opportunity costs of time, which are excluded from conventional definitions of self-control failure). But why would patience be costly? This perspective is especially puzzling in light of the large number of examples of evolved patience. For example, a male rhesus macaque can wait months to gain weight in preparation for the fights associated with mating season (Strier, 2018; Rawlins and Kessler, 1986). There is no obvious reason to think these apparently lazy males are exerting months of difficult self-control.

**Evaluating some theories about why self-control fails**

Perhaps the most influential explanation for self-control failure is the idea that control relies on a limited internal resource (Baumeister and Heatherton, 1996; Muraven and Baumeister, 2000; Baumeister et al., 1998; Inzlicht and Schmeichel, 2012). In the influential *strength model*, or ego depletion model, self-control is demanding in the same way that muscular movement is (Baumeister et al., 2007; Baumeister et al., 1994). That is, self-control requires effort, it depletes some central reserve, and improves with practice (Muraven et al., 1998; Baumeister et al., 98; Baumeister et al., 2006). Glucose was proposed as this energy store; perhaps self-control requires mental activity that is metabolically costly (Baumeister et al., 2006; Gailliot and Baumeister, 2007; Dvorak and
Simons, 2009). This theory has a natural evolutionary explanation because energy is an obvious limiting factor for any organism.

One limitation of this view is that there is no obvious neuroscientific reason why self-control is metabolically costly. Another is that the strength theory is not empirically supported. Meta-analyses and large replications indicate that small study bias and publication bias likely led to over-inflated estimates of depletion effects, which may not exist at all (Carter et al., 2015; Hagger et al., 2016; Carter and McCullough, 2013; Xu et al., 2014; Inzlicht and Berkman, 2015). Likewise, the idea that glucose serves as the reservoir has been successfully challenged (Carter and McCullough, 2013; Molden et al., 2012; Vadillo et al., 2016). This is not to say that self-control does not vary systematically, or flag with fatigue, just that ego depletion cannot account for most of its effects (Muller and Apps, 2018). In any case, the prominent failures of the ego depletion hypothesis are an important motivator for the questions I raise here.

Another idea is that more self-control requires a larger brain (MacLean et al., 2014; Stevens, 2014). Failures, in this view, come from insufficient mental resources – associated with brain volume – and the tradeoffs in self-control are the same as those associated with brain size. Support for this view, for example, comes from a major study comparing self-control in thirty-six species showing that absolute (and not relative) brain size predicted self-control strength across species (MacLean et al., 2014). Likewise, Stevens examined intertemporal choice performance in thirteen primate species (Stevens, 2014; see also Stevens and Muhlhof, 2012). Among other variables, absolute brain size (and again, not relative brain size), predicted self-control. These findings suggest that something about larger brains allows us to wait longer.

However, the specific tasks used in both the studies have been challenged as measures of self-control (Beran, 2015; Fawcett et al., 2012; Stephens and Anderson, 2001; Hayden, 2016). Even if these tasks measure a combination of self-control and other processes, the evidence linking brain size to self-control may instead demonstrate a link to other factors. Indeed, brain size correlates with many other factors that may contribute to preferences in these tasks, such as metabolic rate and lifespan (Stevens, 2014). More
fundamentally, these studies do not offer an explanation of why larger brains would lead to more self-control. While it seems reasonable that larger brains lead to complex mental abilities like general intelligence and social intelligence, self-control would seem to be computationally simple (Tobin and Logue, 1994; Sol et al., 2010; Byrne and Corp, 2004; Deaner et al., 2007). It just requires a computation and comparison of reward rates associated with each option. Animals, even ones with small brains, are highly adept at estimating and maximizing reward rates (Stephens et al., 2007; Stephens and Krebs, 1986). Even bees and ants, which have minimal nervous systems can do it nimbly (Wendt and Czaczkes, 2017).

Yet another explanation for poor self-control has to do with the importance of prospection for self-control (Bulley et al., 2016). Specifically, it has been argued that episodic foresight, the ability to simulate the future and reason about it, is critical (Suddendorf and Corballis, 1997; Buckner and Carroll, 2007). According to this view, failures of self-control result from failures to prospect. This viewpoint has several limitations, however. Most importantly, self-control is more widespread in the animal kingdom than prospection is (MacLean et al., 2014). For example a rat can exhibit self-control but likely has no episodic foresight (Tobin and Logue, 1994). Moreover, while self-control may benefit from prospection, it is not essential (Kwan et al., 2012; Stephens and Stevens, 2008; Craver et al., 2014). Indeed, it may be that prospection is critical for flexible self-control, but not for successful simple self-control itself (Bulley et al., 2016).

Finally, several scholars have focused on the adaptive benefits of poor self-control (Stephens and Anderson, 2001; Kacelnik, 2003). Thus, for example, patience entails both an interruption risk and a collection risk (Fawcett et al., 2012; Stephens, 2002; Kagel et al., 1986; Sozou, 1998). Both of these risks increase the opportunity cost of waiting relative to selecting the immediate option (Stephens, 2002). Likewise, self-control may be fit to environment. For example, marmosets and tamarins have diverged relatively recently but have very different ecological niches. Insectivorous tamarins move quickly to catch their prey and discount time steeply (and thus ostensibly have poor self-control); marmosets, which specialize in tree-sap exudate, need to be patient to feed and are so, but
they discount space steeply (but, by conventional definitions, have good self-control). This approach can explain variations in self-control observed across species, including primates (Rosati et al., 2007; Stevens et al., 2005; Stevens et al., 2005; Heilbronner et al., 2008).

This perspective is valid, and is likely at least partially correct (but see Henly et al., 2007), but it does not help us with the evolutionary puzzle that interests us here. That is, if a smaller-sooner reward offers a larger expected rate of intake, then choosing it is, by my definitions here, is good – not poor - self-control. In other words, if an idealized perfectly self-controlled decision-maker would make the same choices, we cannot – based on behavior (which is all we can measure in animals) - call it a self-control failure. And in the case that self-control inarguably fails, as in a macaque that chooses smaller/sooner options more than interruption/collection risk indicates, the adaptive fit theory does not provide any explanation.

The intertemporal choice task

The intertemporal choice task has long occupied a central place in the self-control research program (Rachlin, 2000; Ainslie, 1974; Berns et al., 2007; Kacelnik and Bateson, 1996; Mazur, 1978; Kirby and Marakovic, 1996; Kurzban et al., 2013, Frederick et al., 2002). This task, which is widely used in both humans and non-human animals, involves a series of choices between options that differ in delay and magnitude. Successful self-control is defined as selection of a larger option with a greater delay or effort cost over a smaller but cheaper/sooner one. Non-human animals generally show discount factors with a half-life (i.e. discount factor, $k$) of a few seconds; humans show a wider range, from these short timescales to factors in the range of weeks to months, but are still impulsive (Green et al., 1994; Kirby and Herrnstein, 1995; Frederick et al, 2002).

The high discount factors typically observed in the intertemporal choice task strongly violate the principles of adaptedness (Stephens and Anderson, 2001; Stephens et al., 2004; Kacelnik, 2003; Pavlic and Passino, 2010; Pearson et al., 2010). That is, animal decision-makers that discount on the order of seconds could not possibly negotiate simple
tradeoffs necessary to survive in the world. For a monkey with a very low (i.e., patient) discount rate of $k=0.05$, the subjective value of an option that is only ~20 seconds away would be reduced by half of its true value. An option that requires two minutes to obtain would have essentially no value. This animal obviously could not make good decisions and survive outside the lab. Indeed, strong arguments have been made that the intertemporal choice task is different in key ways from tasks animals are likely to have faced in their evolved histories (Fawcett et al., 2012; Stephens et al., 2004; Kacelnik and Bateson, 1996; Namboodiri et al., 2014; Blanchard et al., 2013). Measures of time preferences in more naturalistic tasks produce order-of-magnitude improvements in measured self-control; that is, animals seem to have better self-control if it’s measured differently (Schweighofer et al., 2006; Blanchard et al., 2013, 2014, 2015; Blanchard and Hayden, 2015; Pearson et al., 2010; Stephens and Anderson, 2001; Stephens et al., 2004; Stephens and Stevens, 2008).

At a minimum, these results challenge the external validity of the form of self-control measured by the intertemporal choice task (Hayden, 2016). Why would the task lack validity? Even decision-makers with perfect self-control will show apparent self-control failures if they misunderstand the task (Stephens and Anderson, 2001; Stephens et al., 2004). For example, most implementations of the intertemporal choice task use a post-reward buffering structure to avoid “cheating” strategies of choosing the smaller sooner reward to get to the next trial sooner. But this stratagem only works if the animal fully understands the structure of the task. Failure to correctly understand the buffering structure will produce apparently poor self-control in a maximizing forager (Blanchard et al., 2013; Hayden, 2016). Most animals likely either misunderstand or misapply this element of the task (e.g. Stephens and Anderson, 2001; Mazur et al., 1978; Logue et al., 1985; Green et al., 1981; Fox et al., 2013).

There is a second problem that limits the interpretability of the intertemporal choice task. The traditional definition of poor self-control holds, in essence, that poorly controlled decision-makers will over-weight time relative to reward. But decision-makers with poor self-control may more readily over-weight the reward dimension relative to the
time dimension (Paglieri, 2013; Bramlett, 2012; Beran, 2015). The presence of an option that produces a large amount of food is a strong tempter (Boysen and Berntson, 1995). Indeed, the temptation to seek food and ignore costs would seem to be archetypical self-control problem. Decision-makers who succumb to the temptation to choose the larger amount will, in a typical intertemporal choice task, have – by conventional definitions – a surfeit of self-control (Adessi, 2011).

The human intertemporal choice task is not usually implemented with adjusting buffers, and does not usually use primary rewards like food (but see McClure et al., 2007; Hayden et al., 2007; Pearson et al., 2010). As such it does not have the same problems the animal version does. Nonetheless, the relevance of the task to human self-control has been questioned. First, the external validity of the task is quite low, compared to other self-control tasks such as the BART task (Lejuez et al., 2002). Second, humans exhibit several anomalies that cannot be explained through the principle of discounting (reviewed in Frederick et al., 2002). For example, in many cases, humans and animals prefer sequences of rewards in which value increases over time to sequences in which value declines (Frederick et al., 2002; Hsee et al., 1991; Varey et al., 1992; Chapman, 2000; Blanchard et al., 2014). Indeed, it appears that humans preferentially use heuristic strategies that result in discounting-like behavior without discounting (Scholten and Read, 2010).

Towards a cognitive control-based theory: three examples

In moving towards considering the evolutionary causes of poor self-control, is it helpful to begin with a few concrete examples. The first comes in the form of drugs and alcohol. These are common sources of self-control problems, affect a large number of people, and can have deadly consequences. They are clearly ego-dystonic in many cases, and clearly resist even very serious and costly efforts to abstain. So, why have we not evolved the ability to resist? Notably, most drugs, at least in their potent modern forms, were not present in the environment of evolutionary adaptedness (EEA). Most addictive drugs (i.e. those that work on reward pathways) were first made available within the past
few centuries. The industrial revolution has led to new techniques for purifying and delivering the drugs such that today’s drugs of abuse are more potent and addictive than they have been at any previous point in our evolutionary history. For example, although alcohol was likely fermented several thousand years ago, until recently it had a relatively low alcohol content. So addictive drugs work by taking advantage of brain communication networks that were evolved in an environment without them. From a necessarily slow evolutionary perspective, addictive drugs are simply a very novel danger to which we have not yet evolved a solution. Overcoming the temptation to consume drugs and alcohol, then requires making use of general-purpose cognitive faculties. Drugs, then, constitute something of an edge case – we have not evolved mechanisms to overcome drugs and are forced to use – as an inferior backup - our non-specialized cognitive systems.

A second example comes from dieting. Consider that one of the more successful laboratory paradigms in humans has been the diet choice task, in which poor self-control is defined as choosing the tastier but less healthy item (e.g. Hare et al., 2009). Human food resources underwent major shifts at the time of the agricultural and then industrial revolutions – both too recent to have had major effects on the evolution of cognition. In other words, the diet available in the EEA was sufficiently limited that dieting was probably not necessary. Dieting, then, likely can only be implemented by use of deliberate cognitive resources, which conflict with the canalized and inflexible processes that lead us to seek high-calorie food. Tempting food, then, works in some ways like drugs – we have not evolved specialized mechanisms to deal with it, and must make use of a general cognitive system.

A third example comes from a trio of tasks that are often used in animal studies of self-control. In reverse contingency tasks, animals must point to one of two rewards in order to get the other (Kralik, 2005; Izquierdo and Murray, 2004; Chudasama et al., 2006; Boysen and Berntsen, 1995; Silberberg and Fujita, 1996; Stephens and Anderson, 2000; Genty 2004; Shifferman, 2009; Uher, 2008). This task is quite difficult but trainable in some animals. In accumulation tasks a reward is available at any time but builds up the
longer the animal waits. Gaining a larger reward involves inhibiting the taking of the reward, as that would end the accumulation process (e.g., Anderson et al., 2010; Beran et al., 2002; Evans and Beran, 2007; Pelé et al., 2010; Pelé, et al., 2011). Finally, exchange tasks require an animal to keep a small reward in their possession for a period of time before trading it back to the experimenter for a bigger reward (Dufour et al., 2007; Leonardi et al., 2012). One thing that unites these tasks is that they make use of food and not symbols that represent it. Moreover, to overcome self-control the animal must do something that is normally inimical to food receipt. Thus, they involve overriding low-level programming aimed at maximizing caloric intake by use of deliberative overriding systems.

In these three examples of self-control, animals must use general cognitive mechanisms to perform the controlled action and override strong tendencies. These tendencies may be learned through evolutionary time, as in the innate drive for sweet and fatty foods. Or they may reflect the need to override strongly learned action patterns, as in the case of exchange tasks. In any case, what unites these clear self-control examples is the competition between a general cognitive decision-making system and specialized (either learned or hard-wired) decision-making systems.

**Defining self-control as a form of cognitive control**

Self-control is the result of a conflict that arises when competing desires occur. The co-occurrence of both desires requires arbitration. Failure of self-control occur when one desire – the one inconsistent with long-term goals – wins. The other desire – the one associated with poor self-control – wins because it has been given extra heft in the competition, either by evolution or by learning processes.

The brain uses sensory, visceral, and learned information to guide the adaptive selection of actions (Cisek and Pezzulo, 2016). I refer to this process as sensorimotor transformation. When a certain sensorimotor transformation is common, the brain processes it in a more efficient way. I will refer to this as automatization (Botvinick et al., 2014; Shenhav et al., 2017; Shiffrin and Schneider, 1977; Muraven et al., 2000; Miller
and Cohen, 2001). For example, the first time I follow a route across a new campus, it requires attention and dominates awareness. However, if I walk the same path every day, it rapidly becomes automated in my mind, leaving my awareness free to wander or perform other cognitively demanding activities.

Automatization carries several very useful benefits (Miller and Cohen, 2001; Shenhav et al., 2017). Automatic responses are faster. They are less variable and more accurate. They are less susceptible to interference from outside processes. Automating responses leaves room for cognitive control, which is evidently very limited (see below), to engage in other processes. Thus, while walking, I may successfully get to my office, and even perform other complex automated behaviors, like avoiding collisions with other people, while having full capacity to mentally rehearse an important lecture.

But automatization is a double-edged sword. I can offload the processing of sensorimotor information to specialized sub-computations, but those computations are now less penetrable to modulatory influences. Thus, the efficiency that makes it beneficial means it is inaccessible to unexpected, unusual, or rapidly changing goal states. These are precisely the situations that lead to self-control failure. Failure, then, can result from an automated sensorimotor mapping doing what it is supposed to do – but the context happens to be one where the sensorimotor mapping led to a bad outcome.

Failure may also result from an inability to overcome the automated sensorimotor mapping. That is, the two necessarily compete, and the automated one sometimes wins. (If it couldn’t ever win, then it would not serve its purpose of reducing cognitive load). And the ego dystonia associated with self control failure comes from the fact that the brain contains specialized systems that can recognize and signal failures. The self-control failure is not selected for in the conventional sense, but it selected for indirectly, in the sense that it is the unavoidable price worth paying in exchange for the benefits of automatization. Self-control failures are a by-product of processes that produce more efficient but less flexible decisions (Shenhav et al., 2017).

It is worth noting that this argument applies even if automatization exists on a continuum rather than in two discrete states. If we are interested in two competing
processes, then what matters is which is relatively more automatic; if we are interested in several, then what matters is if the one associated with poor self-control is more automatic. It is also notable that controlled processes can be come automated, and then produce poor control. Consider, for example, a child learning, with difficulty, to read text; a few years later, that child is tested on the Stroop task and has difficulty performing accurately in the high conflict condition.

The processes that automate basic sensorimotor mappings operate on both long and short timescales – that is, both the evolutionary timescale and on the scale of much shorter than the lifetime of an individual decision-maker. Both processes are, for my purposes here, similar. Self-control failure can result from suboptimal responses from either learned or evolved mappings. The only important difference is that evolved responses are likely even more ingrained, more canalized, and less susceptible to changing priorities. They may require even larger exertion from the central executive to overcome. It is intriguing, in this view, that some of the most powerful cases of poor self-control (food and drugs) are ones that reflect an evolutionary, not just learning mismatch (reversal tasks).

If self-control is just about cognitive control, why don’t we just evolve a larger capacity for cognitive control? The answer to this question is not yet determined (Miller and Cohen, 2001; Shenhav et al., 2017; Kurzban et al., 2013). It does seem clear, however, control is quite limited (Shiffrin & Schneider 1977; Shenhav et al., 2017). We have difficult with sustained focus and task-switching (Weissman et al., 2006; Boly et al., 2007; Monsell, 2003; Hayden et al., 2010; Logan et al., 2007). Control seems to be qualitatively different from capacities like oculomotor control or form vision, which are excellent.

One possibility is that there are basic computational principles that limit the capacity of any such complex system. For example, there is good reason to think that the brain, especially the prefrontal cortex, has many properties in common with a certain neural network type known as an attractor networks. These networks can be studied in simulations to give insight into their properties, which may reflect the properties of the
real brain. There is evidence that the number of representations that can be kept separate within such networks is limited (Usher and Cohen, 1999; Shenhav et al., 2017). The limit on the number of representations then may impose a hard limit on processing capacity. Crucially, the problem cannot be solved by devoting more resources: they brain may not be able to increase the number of available representations because shared representations provide a critical benefit in the form of allowing generalization, insight, and novel solutions to problems (Musslick et al., 2016b; Feng et al., 2014; Musslick et al., 2016a).

**Implications**

By this perspective, self-control is “just” a type of economic choice (as argued by Berkman et al., 2017). But it is a special one: it is one in which at (1) least one of the options is associated with an intrinsic bias towards or against it. And that means (2) overcoming that choice requires effort. And (3) failure to do so is both costly and ego-dystonic. These are not features of conventional economic choices. Thus, while I agree with Berkman and colleagues, I think they bypass the most interesting part of self-control, its ability to fail.

Do humans have more self-control than animals? This question, often asked, is poorly specified. The cognitive control framework lets us ask it more precisely. Do humans have a greater ability to let goals and changing task demands influence their choices – and rely less on automated mappings? The answer is likely yes, but we need more studies directly comparing the cognitive control abilities of humans and other animals (e.g. Maclean et al., 2014). These tests will have to rely on an understanding of the foraging frameworks of animals, so that measures can be designed appropriately to allow comparison (Calhoun and Hayden, 2014).

If self-control is a type of cognitive control, this suggests that much of the psychology and neuroscience of cognitive control may have direct benefits in helping us to understand self-control as well. Thus, for example, information about the neuroanatomy and neuronal mechanisms of self-control should be directly testable as
theories of self-control. One important area for future research will be to see whether our understanding of cognitive control failure can shed insight into the mechanisms of self-control failure.

From the cognitive control perspective, poor self-control is the default and good self-control is more likely to require deliberate effort (Metcalf and Mischel, 1999). This does not mean that we will always perceive it as such. Our brains are highly practiced at deploying cognitive control flexibly and adaptively, so the conflict between controlled and automatic processes may not rise to the level of consciousness. Or it may rise to the level of consciousness but may be perceived as effortful, difficult, or just distracting. The phenomenology of self-control is poorly studied, but likely to be an important motivator for future research.

Recent work in the field of self-control and in cognition more broadly has challenged the two-systems view on empirical, theoretical, and neuroscientific grounds (Kable and Glimcher, 2009; Kurzban et al., 2013; Inzlicht and Berkman, 2015; Berkman et al., 2017; McGuire and Kable, 2015; Buckholtz et al., 2015). Nonetheless, taking a cognitive control perspective on self-control suggests that this view has at least a few merits. Specifically, any given self-control decision reflects a competition between what can be thought of as distinct brain processes. (This is true whether these two types of processes are regionally or even anatomically differentiated). That doesn’t mean there are two systems; but there are two tendencies: automatic and controlled. And these types may be relative, not absolute; we may have a spectrum of processes ranging from automatic to controlled. But in many specific situations, there will only be two relevant processes competing. In that sense, one could even meaningfully label them hot and cold.

**Implications for the neuroscience of self-control**

The idea that self-control is a type of cognitive control suggests that we do not have a special self-control system or self-control module in the brain. That is, if self-control is continuous with cognitive control, there is no region whose unique purpose (or even one of its purposes) is to drive self-control. Instead, controlled actions likely make
use of a more general control architecture. This control architecture may be modular or may be distributed (Botvinick and Cohen, 2014; Miller, 2000; Hampshire et al., 2015; Eisenreich et al., 2017). It likely includes dorsal prefrontal regions, such as the dorsolateral prefrontal cortex and dorsal anterior cingulate cortex (Botvinick and Cohen, 2014; Miller, 2000; Ebitz and Hayden, 2016; Heilbronner and Hayden, 2016; Shenhav et al., 2017). This approach implies that neural signatures of self-control will be continuous with signatures of cognitive control. Thus, a critical test for this idea would be to compare networks involved in cognitive control with those involved in self-control. This could be done at both the neuronal level and at larger scales.

This idea also has implications about the relationship between self-control and economic choice. Our proposal, in essence, is that self-control is an economic choice in which one option is intrinsically favored over the other, but the other is more consistent with long-term goals. Standard approaches to neuroeconomics are derived from economics, and often involve binary choices, or choices between two goods or bundles of goods in which neither is default, or a priori favored. In contrast, foraging-inspired models of economic choice take as their starting point the idea that choices are between accepting and rejecting single options (Kolling et al., 2014; Kolling et al., 2016; Hayden, 2018; Hayden and Moreno-Bote, 2018). These models, in turn are inspired by ethological observations about the types of decisions that foragers make in natural environments (Stephens and Krebs, 1986; Kacelnik et al., 2011). One of the key differences between accept-reject decisions and binary choices is that we may have intrinsic tendencies to prefer accepting or rejecting, or, because they are computed differently the two types of choices may have different psychological processes and different neural substrates (Kolling et al., 2016).

This idea in turn relates to the idea of affordance competition (Cisek and Kalaska, 2010; Cisek, 2012). Embodied theories of economic choice, going back to Gibson, emphasize that control of action is the ultimate evolutionary driving force in the brain. As such, it is not surprising that we see signatures of action even in supposedly abstract reward areas (Strait et al., 2016; Yoo et al., 2018). From this perspective, stimuli we
encounter in the environment trigger affordances or plans for potential actions to take. The decision about whether to take that action depends on some thresholding process (whose exact nature remains to be delineated). But it is this process that determines the outcome of most self-control decisions.

**Conclusion**

Self-control is often taken as a given: we have poor self-control, but if we tried harder we would do better. From the psychological perspective, we are flawed. But from the evolutionary perspective, we are descended from a long line of successful foragers, and every element of our psychology has some potential explanation in our evolutionary history. Thus each of our major flaws – our tendency to lower back pain, the weakness of our anterior cruciate ligaments, our tendency to get kidney stones, our inability to fly – demands an explanation. Our poor self-control is a major flaw as well. We can lament it and urge ourselves to do better in the future, but we can also be a bit more objective and ask why poor self-control is so universal. While the phylogenetic origins of self-control failure remain to be worked out, the weight of evidence suggests that the neural origins lie in the domain of cognitive control.
Acknowledgements
This work was supported by NIH R01 DA038615. I thank Amitai Shenhav, Becket Ebitz, and Habiba Azab for comments on a working version of this manuscript.

References


44. Fawcett TW, McNamara JM, Houston AI. 2012 When is it adaptive to be patient? A general framework for evaluating delayed rewards. *Behav. Processes* 89, 128–


122. Shenhav A, Musslick S, Lieder F, Kool W, Griffiths TL, Cohen JD, Botvinick MM. 2017 Toward a Rational and Mechanistic Account of Mental...


139. Strier KB. 2018 Primate Behavioral Ecology. 5th edn. T&F India.


