Introduction

*José Luis Bermúdez**

Self-control raises fundamental issues at the heart of practical decision making, human agency, motivation, and rational choice. Unsurprisingly, therefore, it has been studied and discussed within several different academic literatures. Psychologists, philosophers, and decision theorists have all brought valuable perspectives on and insights into how to model self-control, different mechanisms for achieving and strengthening self-control, and how self-control fits into the overall cognitive and affective economy. Yet these different literatures have remained relatively insulated from each other.

The chapters in this collection bring those literatures and approaches into dialog by focusing on the rationality of self-control. This Introduction begins by comparing and contrasting the different approaches to self-control taken in philosophy, psychology, and decision theory, respectively. After setting up a schematic and illustrative puzzle of self-control as a framework for mapping out the different contributions, I then draw out some principal themes running through the different chapters and briefly introduce each contribution.

Philosophers, psychologists, and decision theorists typically approach the topic of self-control in subtly different ways. These approaches are in many ways complementary, but it will be helpful to begin by sketching out some characteristics of the different disciplinary perspectives on self-control before drilling down more deeply into the themes and arguments of the chapters in this collection. (*Warning:* I will be painting selectively with broad strokes of the brush.)

I.1 Philosophy: The Greek Background

Philosophical engagement with self-control dates back at least as far as Socrates, and subsequent discussion (at least within the Western tradition)

* Work on this Introduction was supported by a grant from the Philosophy and Psychology of Self-Control Project funded by the John Templeton Foundation.

has been very much circumscribed by ways of thinking about self-control directly traceable to Socrates, Plato, and Aristotle.[1] Ancient Greek discussions of self-control were framed by the guiding idea that self-control is a virtue, or at least a character trait, typically defined by contrast with its opposite, weakness of will.

The Greek word *akrasia* that has now become standard terminology for talking about weakness of will only appeared in mainstream philosophical discussion in the work of Aristotle, but the phenomenon itself was discussed extensively by Socrates and Plato. In the dialog *Protagoras*, Socrates counterintuitively and provocatively denied that weakness of will, as standardly construed, exists. The qualification is important. Socrates was not denying the existence of weak-willed behavior. He was not denying that people often have a second glass of wine or a third slice of cake when they think it best that they abstain. His objection was to standard ways of thinking about weakness of will. In particular, he denied that weak-willed behavior comes about when one's knowledge of what one ought to do is somehow overruled by the pleasures of the moment. Knowledge, for Socrates, cannot be "a slave, pushed around by all the other affections."[2] And so, by the same token, self-control cannot be a matter of resisting temptation and, more generally, mastering the emotions.

In the *Republic*, Plato gave an exceptionally clear articulation to precisely the conception of self-control and weakness of will that Socrates had rejected. For Plato, the space for self-control is set by ongoing conflict between the rational and irrational parts of the soul (*psuche*). The irrational parts of the soul dominate in the weak-willed person, while self-control results when the rational part of the soul prevails. Aristotle's condensed and often difficult to understand discussion in the *Nicomachean Ethics* incorporates elements from both the Socratic and the Platonic perspectives but definitely comes out closer to Plato than to Socrates.

In one respect, though, all three of the great Greek philosophers are in agreement. They each take weak-willed behavior to be a paradigm of practical irrationality and self-control to be a rational ideal. For Plato and Aristotle, the practical rationality of self-control is achieved through the

[1] The essays in Bobonich and Destrée (2007) extend the discussion of *akrasia* through the later Greek philosophers up to Plotinus. Outside the Western tradition, self-control has been discussed by classical Indian and Chinese philosophers. In the Nyâya dualist tradition, self-control (*svâtantrya*) was taken by some philosophers to be a distinguishing mark of the conscious (see Chakrabarti 1999: chap. 7), while the control of the emotions was a recurrent theme in early Chinese philosophy, as discussed in Virág (2017).

[2] Plato, *Protagoras*, 352 B.C., trans. W. K. C. Guthrie. In Hamilton and Cairns (1961: 344).

exercise of willpower – a conscious effort to resist temptation and master affects and emotions. For Socrates, in contrast, weak-willed behavior is the result of an agent miscalculating the overall pleasures and pains that will result from a given course of action – and so, correlatively, the practical rationality of self-control comes from correctly applying what he calls the *art of measurement* (correctly assessing the balance of pleasures and pains in every action and situation).[3]

This focus on practical rationality is the hallmark of most subsequent philosophical discussions of self-control. Such a focus is typically aligned with a model of practical reasoning and agency that invokes intentions and judgments about what is all-things-considered better. The focus tends to be on explaining how all-things-considered judgments can either prevail or be overruled. How do we need to think about the human mind and the human motivation system to understand how self-control can take place and how it can fail?[4]

This way of thinking about self-control is often synchronic. That is, the discussion is of how, at a given moment of choice, it is possible for the passions to be mastered, temptation overcome, and better judgment vindicated. But some philosophers, particularly those influenced by discussions in decision theory (on which see further later), have explored how self-control can play out in a diachronic context, where the problem is how an agent can adhere to earlier resolutions and commitments when motivations change, both when such motivations are anticipated and when they are not.[5]

I.2 Psychology: Mechanisms of Weakness and Mechanisms of Control

From the perspective of psychology, the focus has tended to be on mechanisms – both the mechanisms that make self-control necessary (the mechanisms of weakness, as it were) and the mechanisms that make self-control possible (the mechanisms of strength). Considerations of rationality have not typically been at the forefront of discussion, but the emphasis on

[3] Or at least that's how he describes things in *Protagoras*, where he derives this view from a version of psychological hedonism. Scholars disagree about whether the psychological hedonism in *Protagoras* is Socrates' considered view or simply a dialectical tool that suited his purposes at the time. The former interpretation is defended in Irwin (1995: §60, pp. 85–87).

[4] For recent, influential, and representative discussions, see Holton (2009) and Mele (2012).

[5] See, for example, the essays by Michael Bratman collected in Bratman (1999).

finding techniques for enhancing self-control often suggests an implicit assumption that self-control is practically irrational.

Pioneering studies of animals and humans have shown that the need for self-control can be conceptualized in terms of different ways of discounting the future. Very few subjects, human or animal, value goods to the same degree irrespective of the time that they will be received. Typically, the value a good is perceived to have decreases more the longer the time before it is received. The rate at which value diminishes with temporal distance is a function of how the subject discounts the future. This is particularly relevant to a class of cases that will be discussed further later, namely where good intentions are thwarted by temptation. What happens when agents backtrack on commitments (breaking a diet, for example, or not following through on an exercise program or a savings plan)? When the plan to diet, to exercise, or to save is made, the perceived long-term benefits of following it outweigh the anticipated, but still distant, fleeting rewards of backsliding. And yet, when those fleeting rewards (the extra cake, the lie-in, or the extra disposable income) are near at hand, temptation can overcome the best resolutions. How should this kind of preference reversal be understood?

Influential studies by George Ainslie, Howard Rachlin, and others have shown that psychological phenomena such as these can be understood in terms of discount functions that have a particular form. These are discount functions where the rate of change varies over time. In what are known as *exponential discount functions*, the discount rate remains constant. This means that a given delay, say a day, will be accorded the same weight whenever it occurs. So, for example, if I prefer $10 today to $11 tomorrow and I discount the future in an exponential manner, then I will prefer $10 in 100 days to $11 in 101 days. Most people do not take this approach, however. Even if I am not prepared to wait until tomorrow for an extra dollar, the same reward would probably lead me to extend my wait from 100 days to 101 days. This pattern of preferences can be understood in terms of discount functions that are *hyperbolic*. In a hyperbolic discount function, the rate of discounting is affected by the (temporal) proximity of the item being discounted. As we will see in more detail later, this means that agents with hyperbolic discount functions can succumb to preference reversals. Psychologists have explored the relation between discounting and various types of addictive and compulsive behavior.

Preference reversals are not inevitable. Temptation does not always win out. But what are the mechanisms that make this possible? Outside the laboratory, people often talk about self-control in terms of the exercise of

willpower, often treated as a kind of psychic force that some people have more than others and that can be cultivated and strengthened. This basic idea goes back at least as far as Freud, but one influential development within scientific psychology of this very intuitive idea is the ego depletion theory, originally due to Roy Baumeister and Dianne Tice.[6] According to ego depletion theory, willpower is a limited resource than can be used up and run down completely. In the influential experiment that launched the theory, students who had held back from freshly baked chocolate chip cookies and instead snacked on radishes gave up much sooner on a tricky geometric puzzle than either a control group or students who had eaten the cookies. As with a number of areas of social psychology, however, doubts have been raised about the experimental support for ego depletion theory.[7]

Ego depletion theory makes much of the image of self-control as being like a muscle both in so far as it can be used to exhaustion and in so far as it can be strengthened through use. This second idea is independent of the first, and a number of experiments seem to show that regular "exercise" can strengthen willpower. So, for example, Megan Oaten and Ken Cheng found that subjects who pursued a two-month physical exercise regime did better than control subjects on standard laboratory self-control tasks.[8] Other possibilities explored for improving the mechanisms of self-control include implementation intentions (developing determinate strategies in advance for dealing with specific temptations) and defusing temptation by representing the object of temptation in a "cool" rather than "hot" motivationally charged way (or, alternatively, representing the long-term gain in a hot rather than cool way).[9]

I.3 Decision Theory: Problems of Dynamic Choice

Classical decision theory codifies the conception of instrumental rationality dominant in the social sciences, most prominently in economics (excluding subfields of economics such as behavioral economics and experimental economics, which explicitly explore alternatives to standard models of rationality). Decision theorists typically model instrumental

[6] See Baumeister et al. (1998), and for a more popular presentation, see Baumeister and Tierney (2011).
[7] A multilaboratory replication project sponsored by the Association for Psychological Science found little evidence for the ego-depletion effect (Hagger et al. 2016).
[8] Oaten and Cheng (2006). Relatedly, see Muraven et al. (1999) and Muraven (2010).
[9] For a review of research into implementation intentions, see Gollwitzer and Sheeran (2006). Walter Mischel originated the hot/cool systems approach. See, for example, Mischel and Ayduk (2004) and Mischel et al. (2011).

rationality (for individual decision makers in nonstrategic situations) through some form of expected utility theory. Popular versions of expected utility theory all make the same basic prescription, which is that a rational decision maker will choose an option that maximizes utility when the utility assigned to an option's outcomes is appropriately weighted by the probability that it will occur. In other words, rational decision makers maximize expected utility.

From the perspective of classical decision theory, self-control presents something of a puzzle. This is because exercises of self-control are typically acts (or omissions – when self-control leads me not to act) that refer back to an antecedent decision or commitment. So, for example, my staying up late preparing my class is an exercise of self-control by virtue of my commitment to being adequately prepared for my class. But for this to count as an act of self-control, it would seem that the agent's prior decision or commitment cannot motivationally outweigh the agent's current desires and preferences. If my desire to be the designated driver is stronger than my desire for a drink, then how am I exercising self-control in declining the drink? Surely, I am just doing what I most want to do. So, in instances of self-control, the agent's current desires and preferences will typically outweigh her prior decision or commitment.

This is problematic for classical decision-theoretic accounts of dynamic choice (sequences of choices) because a rational decision maker will only take into account her utility assignments at the moment of choice, ignoring any earlier assignments not reflected in her current assignments. This is often called the *historical separability of preferences* (or the *time separability of preferences*).[10] So, if the agent has undergone a preference reversal (so that the fleeting temptation has become more attractive than the long-term goal), the separability of preferences seems to make completely irrelevant the high utility previously assigned to reaching the long-term goal. On one interpretation of the time separability of preferences, classical decision theory prescribes that rational decision makers should choose *myopically*, looking only at the here and now and ignoring how they have valued things in the past.

Against that prescription some decision theorists have proposed that rational agents in such a situation need either to be *sophisticated choosers* or *resolute choosers*.[11] A sophisticated chooser in effect "ties herself to the mast," like Odysseus preparing himself to sail near the Sirens. A sophisticated chooser

[10] See McClennen (1990) for a comprehensive and influential discussion of separability assumptions in decision theory.

[11] Sophisticated choice strategies originate with Strotz (1956) but the terminology with Hammond (1976).

might give her car keys to a friend, for example, or agree to forfeit money if she misses a workout. Such precommitment strategies are intended effectively to remove the option of succumbing to temptation. The sophisticated chooser does not exercise self-control at the moment of choice. She exercises it in advance, reasoning backward in a way that eliminates what she believes (anticipating her future preferences) to be nonfeasible options. Sophisticated choice remains consistent with the time separability of preferences because the sophisticated chooser looks forward but reasons backward.

A resolute chooser, in contrast, eschews the precommitment strategies of sophisticated choosers.[12] He sticks to his guns at the moment of choice, conforming to his earlier plan even in the face of temporarily reversed preferences. The resolute chooser has nonseparable preferences because he is swayed by previous valuations that are motivationally outweighed at the moment of choice. Because the time separability of preferences is a natural extension of expected utility theory when it is applied in a dynamical context, this means that the resolute chooser will not be an expected utility maximizer.[13]

The concept of resolute choice is the closest that decision theory comes explicitly to modeling self-control. And yet it raises two fundamental questions. The first is how (if at all) the rationality of resolute choice can be defended within the instrumental perspective of decision theory. There is at least a prima facie tension between instrumental models of rationality and the everyday phenomenon of self-control. The second is how resolute choice is even possible. The discussion in the decision-theory literature has focused primarily on the rationality of resolute choice. The actual mechanisms have received little to no discussion in that literature.

These two questions relate closely to the issues respectively explored in the preceding two sections. Clearly, ongoing discussions in philosophy, psychology, and decision theory explore intersecting and overlapping topics. The chapters in this volume break ground in drawing those discussions together. I turn now to setting up a schematic puzzle of self-control that will allow me to introduce some of the principal themes of the individual contributions.

I.4 A Paradigm Case of Self-Control

Drawing on some of the threads emerging in the preceding section, I will define a paradigm case where self-control seems to be required. This case

[12] For a defense of resolute choice, see McClennen (1990, 1998). See also Gauthier (1997).
[13] See McClennen (1990: §7.5, esp. n. 7).

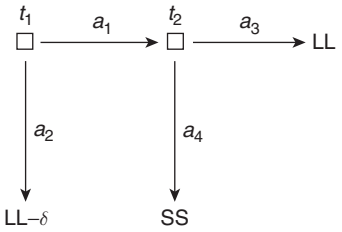

Figure I.1 The paradigm case of self-control represented as a sequential choice problem. The moment of planning is at time t_1 with the moment of choice at time t_2. At t_1 the agent has a choice between making a precommitment to LL (which would guarantee receiving LL, but at a cost, namely δ) or continuing to t_2. At t_2 the choice is between SS and LL.

raises some fundamental questions about the nature, exercise, and rationality of self-control. I will then go on to situate the chapters in this volume through the different answers that they offer to these questions.

The paradigm case that I am proposing is similar to standard cases discussed in the psychological and decision-theoretic literatures. Let us assume that an agent makes at time t_1 a commitment or resolution to pursue a large, long-term benefit at a later time t_3. At a time t_2, later than t_1 and earlier than t_3, the agent has the opportunity of abandoning the long-term commitment in favor of a small, short-term reward. I follow standard practice of using the abbreviations LL (for *larger, later*) and SS (for *smaller, sooner*). At the time of making the resolution, the (discounted) value of LL is more powerfully motivating than the (discounted) value of SS. That is to say, the utility that the agent assigns at time t_1 to the future receipt of LL is greater than the utility she assigns to the future receipt of SS. However, by time t_2 the agent's preferences have (temporarily) reversed, and now SS motivationally outweighs LL. Because t_2 is the moment of choice, this is an opportunity for the agent either to exercise self-control or to succumb to temptation and weakness of will.

We can depict this paradigm case as a sequential choice problem, as illustrated in Figure I.1. In addition to the option at time t_2 of exercising self-control and holding out for LL instead of succumbing to SS, Figure I.1 represents the sophisticated choice option of adopting some sort of precommitment strategy. The outcome of the sophisticated choice option (a_2) is the long-term reward LL minus the cost of precommitment, represented by δ. So, to complete the dynamic choice typology sketched out here, option a_4 represents myopic choice (weakness of will or succumbing to temptation), while option a_3 is the self-controlled, resolute choice.

This schematic decision problem raises three distinguishable, but definitely interrelated, sets of questions. The first set of questions clusters around the concept of rationality. Most obviously, one might ask how a rational agent should tackle this decision problem. Most of the contributors to this volume are working within (or at least exploring the consequences of) a broadly instrumental decision-theoretic conception of rationality, so they typically approach the issue by thinking about how, if at all, instrumental theories of rational choice tackle decision problems, such as our paradigm case, that seem to require the exercise of self-control. The chapters by Thoma (Chapter 1), Peterson and Vallentyne (Chapter 2), and Weirich (Chapter 3) are all situated within this general area, exploring the claims of decision theory to provide a standard of rationality for decision problems of the basic type of the paradigm case just sketched out (often extending the discussion to more complicated decision problems).

A second set of questions clusters around the mechanisms responsible for the basic preference reversal that gives the decision problem its force. How does a space for temptation arise, even in the face of a strong resolution? What can we learn about self-control and how to exercise it from studying what makes weakness of will possible? And, moreover, issues of rationality arise here too. Is it really the case, as is often assumed, that the psychological phenomena that create a space for weakness of will betoken a degree of practical irrationality? Questions such as these come to the fore in the chapters by Ahmed (Chapter 4), Green and Myerson (Chapter 5), Rachlin (Chapter 6), and Andreou (Chapter 7).

Finally, a third set of questions clusters around the mechanisms that potentially lead to self-control – to taking option a_3 and holding out for LL rather than choosing myopically and opting for a_4 and SS. Even if you think that a rational agent will, all other things being equal, hold fast to commitments in the face of temptation (as opposed to either giving into temptation or adopting a sophisticated precommitment strategy), the question still arises as to how that can actually happen, given the agent's motivational profile at time t_2, the moment of choice. The chapters by Bermúdez (Chapter 8), Mele (Chapter 9), Gold (Chapter 10), and Easwaran and Stern (Chapter 11) present a range of different perspectives on this important question.

The remainder of this Introduction uses this general mapping of the theoretical landscape to introduce the individual contributions.

I.5 Rationality, Dynamic Choice, and Self-Control

Can it be rational to exercise self-control to resist temptation? As suggested earlier, this is a puzzling question. On the one hand, agents seem to do better in the long run and by their own lights if they do exercise self-control – which suggests that self-control is instrumentally rational. On the other hand, though, self-control requires overriding one's current desires and for that very reason seems to be instrumentally irrational. Chapters 1 through 3 present different perspectives on this puzzle.

In "Temptation and Preference-Based Instrumental Rationality" (Chapter 1), Johanna Thoma evaluates two of the most frequently canvassed lines of argument in this area. The first line of argument employs a two-tiered strategy, with exercises of self-control counting as instrumentally rational when the level of evaluation is shifted from individual actions to deliberative strategies. David Gauthier's analysis of a sequential prisoner's dilemma modeled on Hume's famous example of the two farmers at harvest time is a good example of how this might work.[14] But, according to Thoma, all such strategies are doomed to fail because the shifted preferences that are problematic at the level of individual actions simply reappear at the higher level of deliberative strategies (because strategies that permit selective exceptions to accommodate temptation will typically be preferred to strategies that do not).

The second line of argument is exemplified by Ned McClennen's discussion of resolute choice (see footnote 12). What makes resolute (i.e., self-controlled) choice instrumentally rational, he argues, is that the resulting plan is Pareto superior to its sophisticated and myopic alternatives when evaluated over the agent's successive selves (time slices) – in other words, it improves things for some time slices without leaving any time slices worse off. The problem with this approach, Thoma maintains, is that it imposes irreconcilable demands on successive time slices. On the one hand, they must be sufficiently unified to care about each other's preferences and resolutions, but on the other, they must be sufficiently independent of each other that the preferences of the current time slice do not immediately trump those of the other time slices. Thoma sees no way of combining these demands.

Both lines of argument share a common assumption about rationality, namely that the standard for assessing how instrumentally rational a given

[14] See Hume (1739/1978: III.2.5, 520–21) for the original example and Gauthier (1994) for the two-tiered strategy.