**Is Synchronic Self-Control Possible?[[1]](#footnote-1)**

**Abstract:** An agent exercises instrumental rationality to the degree that she adopts appropriate means to achieving her ends. Adopting appropriate means to achieving one’s ends can, in turn, involve overcoming one’s strongest desires, that is, it can involve exercising *synchronic self-control*. However, contra prominent approaches, I deny that synchronic self-control is possible. Specifically, I draw on computational models and empirical evidence from cognitive neuroscience to describe a naturalistic, multi-system model of the mind. On this model, synchronic self-control is impossible. Must we, then, give up on a meaningful conception of instrumental rationality? No. A multi-system view still permits something *like* synchronic self-control: an agent can control her *very strong* desires. Adopting a multi-system model of the mind thus places limitations on our conceptions of instrumental rationality, without requiring that we abandon the notion altogether.

**Keywords:** Instrumental rationality, synchronic self-control, folk psychology, naturalism

**1. Introduction**

The possibility of synchronic self-control – an agent’s ability to act contrary to her strongest desires – rests on an apparent inconsistency. On the one hand, a widely-held philosophical truism about motivation implies that individuals always act on their strongest desires. On the other, everyday experience suggests that people can exercise in-the-moment self-control and hence do not do always act on these desires. The challenge of reconciling the truism and the experience amounts to a puzzle in the philosophy of action, not unlike the puzzle regarding the possibility of weakness of will (Orlandi and Stroud, *ms*).

To see the precise nature of the puzzle, we can analyze Kennett and Smith’s (1996, 1997) characterization of the issue. According to the philosophical truism:

1. If an agent *S* most strongly desires to perform some action *A*, and if she believes herself free to *A*, then she will try to do *A*, if she does anything at all intentionally.[[2]](#footnote-2)

And synchronic self-control involves an agent’s regulating her strongestdesire. To capture this feature, then, *S* exercises synchronic self-control during some interval *I* with respect to some action *A* just in case:

1. During *I*, *S* most strongly desires to perform *A*, and is free to do *A*.
2. Nevertheless, *S* intentionally doesn’t do *A* during *I*.

Hence, according to (1) and (2), (3) is impossible.

There are three prominent responses to the foregoing puzzle of synchronic self-control: Jeanette Kennett and Michael Smith’s ‘non-actional’ account (1996, 1997), Alfred Mele’s ‘ancillary action’ account, and Chandra Sripada’s ‘multi-system’ account. Each of these theories aims to establish and explain the possibility of synchronic self-control. In doing so, each makes claims about our inner motivational machinery, and each has implications for establishing the nature and limits of instrumental rationality. The stronger an agent’s capacity to act in accordance with her considered judgments, in spite of her strongest desire to do otherwise – that is, the stronger her capacity for synchronic self-control – the stronger her capacity for instrumental rationality. Philosophers thus have strong normative motivations for securing the possibility of synchronic self-control (Section 2).

Problematically, it is difficult to adjudicate between these or other solutions to the puzzle of synchronic self-control. Like most analyses in the philosophy of action, discussions of synchronic self-control are undertaken using philosophical folk psychology.But philosophical folk psychology’s methodological commitments are overly permissive, and so leave the debate vulnerable to intractable disagreements (Section 3).

These are not grounds for rejecting philosophical folk psychology outright. There are pragmatic arguments for subscribing to philosophical folk psychology as a theory of action, and for conducting research in light of its extant methodological commitments. There are, however, good reasons to make amendments to philosophical folk psychology. To guide such amendments, philosophical folk psychology must be supplemented with an added, naturalistic constraint. Proponents of philosophical folk psychology can adopt *minimal*, *moderate*, or *comprehensive* versions of this naturalistic constraint (Section 4).

With this added methodological commitment in place, it becomes clear that Kennett and Smith’s and Mele’s accounts both fail to satisfy even the *minimal* version of the naturalistic constraint. These views assume a *single-system* model of the mind. But our best science shows that *multi-system* models of the mind are our best candidates for explaining deliberation, decision-making, and action*.* Sripada’s view has an advantage over its two nearest competitors in that it assumes a multi-system model of the mind. But it fails to meet the *comprehensive* naturalistic constraint, because it omits a principle for arbitrating between systems, a necessary feature for this type of account (Section 5).

Drawing on computational models of decision-making and empirical theories of mind, I describe a novel multi-system model of the mind that applies a widely-accepted, ‘winner-takes-all’ principle of arbitration. The view is in line with our best available computational, neuroscientific, and psychological evidence about choice and action. The upshot of this view, however, is that this principle of arbitration precludes synchronic self-control in the strong sense characterized above. Although agents can have all manner of *strong* desires, once the process of arbitration is complete – barring accidents of implementation – agents act on their *strongest* ones (Section 6).

Full-fledged synchronic self-control is thus at present incompatible with the comprehensive naturalistic constraint. Still, weaker forms of something *like* synchronic self-control are possible. An agent can control her *very strong* desires. In fact, an agent can control her very strong desires most successfully by understanding how her own motivational architecture works and, specifically, by leveraging the ‘winner-takes-all’ arbitration system underpinning her choices and actions. I consider the implications of this conclusion for the notion of instrumental rationality (Section 7).

**2. Three views of synchronic self-control**

An agent exercises instrumental rationality to the degree that she adopts an appropriate means to achieving her end. To use an example from Kennett and Smith (1996), Frog and Toad exercise instrumental rationality to the degree that they avoid eating a cookie in order to maintain their health. But adopting appropriate means can involve overcoming one’s very strong or even one’s *strongest* desires, where these desires can be directly at odds with one’s intended ends. In such cases, adopting appropriate means can directly depend on an agent’s relative capacity to exercise synchronic self-control.

The three prominent philosophical treatments of the issue agree that, despite the puzzling nature of synchronic self-control, the phenomenon remains possible. The general strategy vis-à-vis the puzzle as characterized above is to re-describe (1), and thereby show that (1\*-3) can be made consistent. Nonetheless, these views disagree on the degree to which synchronic self-control can be exercised. These views thus offer differing characterizations of adopting appropriate means, i.e., of instrumental rationality and its intrinsic limitations.

***2.1 The ‘non-actional’ account***

Of the three views, Kennett and Smith (1996, 1997) defend the most circumscribed version of synchronic self-control. They argue that passive, i.e., non-actional thoughts modify desires and actions in such a way as to make synchronic self-control possible.

Kennett and Smith’s conception of action relies on a theory of intrinsic and extrinsic desires. Intrinsic desires are sought for their own sakes: common examples of intrinsic desires include a desire for health or a desire for pleasure. Intrinsic desires are necessary but not sufficient for action, however: they must be combined with occurrent beliefs in order to elicit action. Specifically, occurrent beliefs permit the transmission of intrinsic desires across a means-end relation, making corresponding actions possible. For instance, an individual may have an intrinsic desire to experience pleasure and an occurrent belief that eating a peach would bring about a pleasurable experience. If she is instrumentally rational, she will form an extrinsic desire to eat a peach, where this extrinsic desire is the combination of the intrinsic desire and the occurrent belief. The occurrent belief will transmit her intrinsic desire for to experience pleasure across the means-ends relation, making it possible for her to eat the peach.

Building on this framework, Kennett and Smith’s thinking about SSC can be captured as follows:

1. At *t1*, Frog and Toad are presented with cookies. They have an intrinsic desire for health and an intrinsic desire for immediate pleasure. Their strongest extrinsic desire is to eat the cookies.
2. Also at *t1,* Frog and Toad are disposed to certain thoughts about these cookies. For example, they may think of the cookies as unappealing lumps of fat.
3. Combined with their intrinsic desire for health, having these thoughts produces an extrinsic desire not to eat more cookies. In other words, having these thoughts “enables their intrinsic desire for health to transmit its force across the means-end relation” (1996, p. 69).
4. Consequently, having these thoughts enables Frog and Toad to avoid eating the cookies, and so to exercise SSC.

While Frog and Toad’s intrinsic desire for health remains stronger than their intrinsic desire for pleasure, the extrinsic desire to eat the cookies remains stronger than the extrinsic desire not to eat them (see also Kennett and Smith 1997, p. 70). Still, having the unappetizing thoughts enables Frog and Toad’s desire for health to “transmit its force across the means-end relation” (1996, p. 69). By contrast, the weaker intrinsic desire for pleasure and the associated, stronger extrinsic desire to eat the cookies remain disabled, regardless of relative strength. Kennett and Smith argue that Frog and Toad’s unappetizing thoughts are the products of “reliable cognitive disposition they possess to have such thoughts at such times” (1996, pp. 69-70). As these thoughts are not the products of some desire and means-end belief, they are consistent with Frog and Toad’s other desires remaining strongest and, hence, with their exercise of synchronic self-control.

Kennett and Smith thus defend the possibility of synchronic self-control. At the same time, they offer a relatively circumscribed picture of what synchronic self-control can look like. There must be a kind of fit between an agent’s circumstances and her cognitive dispositions, and these cognitive dispositions must presumably be developed in advance, well before the moment of temptation strikes; even then, an agent must wait for these dispositions to produce an appropriate set of responses. Perhaps unsurprisingly, the passivity of this account has been the object of criticism. Kennett and Smith themselves describe it as “contingent,” and Sripada (2012) has characterized it as a “non-actional” form of synchronic self-control. Sripada argues:

The forms of synchronic self-control that are picked out by the folk term “willpower” appear to be conceived in distinctively actional terms. That is, we normally think that the exercise of willpower is something that we *intentionally* bring about. A person typically *chooses* to exert willpower and such choices, because they are taken to be choices, are often held to be deeply indicative of what is important to the person. Moreover, willpower is accompanied by a distinctive actional phenomenology…. [namely,] feelings of *actively striving to do something* (2012, p. 6)*.*

Sripada argues that Kennett and Smith’s passive, dispositional account neither identifies the kind of active, intentional mechanism for action typically associated with synchronic self-control, nor successfully accounts for the phenomenological experience of action and effort commonly associated with exercising self-control.

Perhaps even more problematically, however, the non-actional account cannot guarantee a robust form of instrumental rationality. If an agent must depend on a set of contingent thoughts that are, in turn, dependent on cognitive dispositions, in order to achieve her ends, then she cannot be said to be ‘adopting appropriate means’ in the fullest sense of the phrase.

***2.2 The ‘ancillary’ account***

Mele (1987, 1992, 1997, 1998, 2003, 2014) offers a more active solution to the puzzle of synchronic self-control. Synchronic self-control is made possible, Mele proposes, in virtue of orthogonal, non-competing desires that can be harnessed in the form of ancillary *self-commands*.

Preliminarily, Mele’s analysis of synchronic self-control can be captured as follows:

At *t1*, Ian is watching television. Ian’s desire to continue watching television (*Dtv*) is stronger than his desire to get back to work (*Dw*).

Also at *t1*, however, Ian is motivated by *Dw* to utter the self-command: "Get off your butt, Ian, and paint that shed!"

Since Ian is in the habit of obeying his self-commands, at *t2*, Ian gets back to work (1987, pp. 69-72).

The central feature of Mele’s account rests on Ian’s ability to utter the self-command. Specifically, Mele must explain how, if Ian’s desire to issue the self-command (*Dc*) is motivated by *Dw*, *Dw* can stillremain weaker than *Dtv*. Mele argues that the motivational force of a given desire depends on three determinants: its positive motivational base, its negative motivational base, and additional cognitive elements. The positive motivational base of a desire refers to “the collection of all occurrent motivations of the agent that make a positive contribution to the motivational strength of that desire” (1987, p. 67). The negative motivational base of a desire refers to “the collection of all occurrent motivations of the agent that make a negative contribution to the motivational strength of that desire” (1987, p. 68).

Taken together, the positive and negative motivational bases constitute a given desire’s total motivational base. But the total motivational base of a given desire doesn’t necessarily determine a desire’s overall influence in initiating a corresponding action. *Additional cognitive elements* can also increase or decrease this motivational influence. Additional cognitive elements include factors such as vividness of representation and increased attention (1987, p. 69). For example, the total motivational base of Jack’s desire to eat a bacon cheeseburger can remain constant, even as the smell of bacon enhances its (the desire to eat the cheeseburger’s) vividness of representation. In turn, this vividness of representation increases Jack’s motivation to eat the bacon cheeseburger.

Mele’s solution to SSC rests on the motivational influence of these additional cognitive elements. In Ian’s case, *Dc* shares the same total motivational base as *Dtv,* meaning that the sum of the negative occurrent motivations discouraging him from work also discourage him from uttering the self-command. Conversely, the negative motivational base of *Dtv* and, by extension, *Dc*, is precisely the positive motivational base of *Dtv*. Ian thus doesn’t want to get up and paint the shed precisely for the same reasons that he wants to watch television instead. But it is *easier*, Mele argues, for Ian to utter the self-command than it is for him to get up and work; because although *Dw* competes with *Dtv* directly, *Dc* only competes with *Dtv* indirectly. *Dc* only competes directly with *D~c*. Moreover, additional cognitive elements can help make *Dc* motivationally stronger than *D~c*. Mele explains (1996, p. 71):

Ian's attention, we may suppose, isfocused much more on the TV than on the prospect of not making an attempt at self-control. Thus, we should not find it surprising that he is more motivated to continue watching TV than he is to refrain from making an attempt at self-control. Moreover, when the thought of uttering the self-command occurs to him, he may be much more attentive to the prospect of surprising his wife with a freshly painted shed than to his reasons for refraining from making the attempt at self-control. This would give *Dc* the advantage of vividness of representation over any desire that Ian may have not to utter the self-command. (In contrast, the thought of getting back to work may make the prospect of continuing to watch TV even more attractive.)

Ian can thus be more motivated to utter the self-command than to not utter it, even when he is more motivated to watch television than he is to work.

Ian’s self-command thus acts as the thin end of the motivational wedge: he may be in the habit of obeying his self-commands; or, alternately, his uttering the self-command may help redirect his attention toward painting the shed, and so provide it with supplementary motivational strength (1987, pp. 71-72). In either event, Ian uses the self-command to initiate what Mele’s calls a *motivational shift*, and thereby to exercise synchronic self-control.

Mele’s view thus explains a subset of the kinds of cases typically captured by folk conceptions of willpower and self-control. Mele’s motivational shift view explains instances of synchronic self-control in which an agent is either doing, or is well on her way to doing, the very tempting activity that she is trying to avoid doing. Ian can issue a self-command because he is already watching television. Sally can issue a self-command *on her* *way to* eating a slice of pie, i.e., even before she starts doing the precise activity she is trying to resist. But neither Ian nor Sally can exercise self-control before they begin watching TV or walking toward the refrigerator. Implementing the motivational shift depends on leveraging the original action, i.e., the television-watching or the walking to the refrigerator containing the slice of pie. As Sripada puts it, Mele’s conception of self-control is a kind of “hanger-on that is done only in addition to the main action(s) that one does on behalf of one’s wayward desire” (2012, p. 9). This feature of Mele’s account leads Sripada to call it the *ancillary* view of synchronic self-control.

Again, Sripada argues, this conception of synchronic self-control it too limited. Folk conceptions of willpower and self-control capture a broader sense of exercising synchronic self-control: “We often think of willpower as a *barrier* that we impose on ourselves over some extended interval of time that blocks us from ever doing (or starting to do) anything in the service of a wayward desire, even though during that entire interval, our wayward desire remains our strongest” (2012, p. 9). For example, Mike may want to run away from a Labrador sniffing his feet for an extended period of time, but successfully avoids doing so by exercising self-control. On Mele’s view, Sripada contends, Mike must either flee the room before issuing the self-command or, at a minimum, begin the process of fleeing by readying to stand, etc.

Although I won’t discuss the details of the issue here, Sripada somewhat overstates the limitations of Mele’s view (for Mele’s own response to this challenge, see Mele (2003, pp. 184-189; 2013, p. 264). The main idea is that Mele can, in some cases, account for an agent’s acting on her strongest desire *without* beginning to act). But Sripada is right to emphasize the problem of temporal duration. Even if Mele’s account could explain how Mike can issue a self-command even before he begins to flee the dog, it cannot explain how Mike “continues to exert willpower to prevent himself from fleeing” for several minutes at a time (2012, pp. 9-10). As its name implies, Mele’s notion of motivational shift depends on an agent’s total motivational base shifting from, say, a relatively weaker positive motivational base to a relatively stronger positive motivational base. Such a shift enables the agent to exercise self-control and, ultimately, causes the moment of temptation to pass. The motivational shift view thus cannot explain how an agent such as Mike can exercise self-control over an extended period of time. Consequently, Sripada concludes, the motivational shift view fails to explain a subset of self-control cases which he refers to as f*ull-blooded* willpower, i.e., those lasting over an “(arbitrarily long) interval *L*” (2012, p. 10). Mele thus presents a more robust version of synchronic self-control and, by extension, instrumental rationality, than Kennett and Smith have done. But his ancillary view nonetheless falls short of guaranteeing unconstrained instrumental rationality, i.e., that kind of instrumental rationality that can both be activated at will and maintained over an extended period of time.

***2.3 The dual-system account***

Finally, Sripada (2012) presents a dual-system approach to the puzzle of synchronic self-control. This approach divides the mind into deliberative and emotional motivational systems. The deliberative system processes reasons and judgments to issue action-desires. The emotional system processes emotions, drives, and cravings to issue action-desires. Each of the two systems can issue relatively independent action-desires: “the two systems can diverge from each other in terms of their motivational upshots, i.e., in terms of the action-desires that arise from the two systems” (2012, p. 12).

Based on this framework, Sripada’s thinking about synchronic self-control can be captured in the following way:

1. Throughout some interval *I*,extending over *t1* . . . *tn*, Sally’s desire to eat a piece of key lime pie, *De,* is her strongest desire.
2. Despite being weaker than *De*, her judgment that it would be best to avoid the pie, *Ja*, initiates a set of regulatory processes, *R*.
3. *R* inhibits *De*.
4. Sallydoes not perform any actions that promote *De*, or start to perform any actions that promote *De,* at any time during *I* (see Sripada, 2012, p. 10).

*R* has a blocking effect on *De*,such that *De* remains Sally’s strongest desire throughout interval *I*,but is nonetheless not acted on.

*R* is specified in two ways. First, *R* is *proprietary* to the deliberative system. Judgments produced by the deliberative system can initiate a set of regulatory processes, *R*. Emotions and desires produced by the emotional system cannot. Second, *R*’s motivational force is *progressive*. The motivational force of *R* cannot be attributed retroactively to the motivational powers of *Ja*. The attribution of motivational force is governed by the principle:

C: The causal powers of processes that operate as a consequence of one’s action cannot be credited ‘backwards’ to the desire that initiated the action (Sripada, 2012, p. 20).

In light of C, *De* remains the strongest desire while still being overcome by *R*.

Taken all together, Sripada’s view uses the features of his dual-system view to defend the possibility of full-blooded synchronic self-control. The proprietary feature ensures that the deliberative system initiates self-control. Sally can act on something other than her strongest desire *De*. The progressive feature guarantees that *Ja* remains weaker than the strongest desire *De*. Throughout *I*, Sally can avoid acting on her strongest desire. Sally can thus exercise an *active* and *extended* form of synchronic self-control and, by extension, exercise a robust kind of instrumental rationality.

**3. Philosophical folk psychology and its limits**

The robust form of synchronic self-control defended in Sripada’s dual-system account is certainly appealing. But it is important to distinguish between the kind of view we would *like* to have – a robust form of synchronic self-control, say, together with a robust version instrumental rationality – from the different kinds of justifications that we provide for such views. The fact that Sripada’s view defends the strongest version of synchronic self-control does not, in itself, mean that the dual-action view offers the *best theory* of synchronic self-control. How can we adjudicate between the competing views and, correspondingly, between the competing pictures of instrumental rationality? I argue that, at present, given the views’ shared, extant meta-theoretical (methodological) commitments, there is no principled way of adjudicating between these views.

The challenge goes something like this. Like most examinations in the philosophy of action and philosophy more broadly construed, the foregoing analyses of the puzzle of synchronic self-control are undertaken using *philosophical folk psychology.* Here, philosophical folk psychology refers to philosophical theories describing human behaviors in terms of mental states such as intentions, beliefs, and desires. These theories are broadly *realist* in nature: they hold that people really experience mental states such as beliefs and desires. They further hold that our everyday descriptions of these mental states are roughly *true*. These analyses precisify and systematize these descriptions to develop full-fledged theoretical accounts of action. These full-fledged theoretical accounts are supported every time they successfully predict human deliberation, choice, and action.[[3]](#footnote-3)

Further, philosophical folk psychology is bound by two methodological commitments. The first, internally-oriented commitment requires that philosophical folk psychological accounts be internally coherent, i.e., logically consistent. We can call this the *internal coherence* constraint. The second, externally-oriented commitment requires that philosophical folk psychological accounts cohere with our everyday introspective experiences (we can refer to these as our intuitions), as well as with our everyday descriptions of these experiences. We can call this the *intuitive plausibility* constraint.

Problematically, these constraints do not provide sufficient grounds for adjudicating between the competing theories of synchronic self-control characterized above. I develop and defend each of these claims in turn.

***3.1 Philosophical folk psychology***

Philosophical folk psychology as characterized above begins with everyday descriptions of mental states and processes and proceeds by *precisifying* and *systematizing* them (for a more rigorous description of this process, see (Lewis, 1972)). The resulting, full-fledged theoretical account is used to *predict* or *explain* action.

We can see a detailed example of this process in Kennett and Smith’s account. First, Kennett and Smith draw on Arnold Lobel’s *Frog and Toad Together* to introduce an everyday description of mental states. In the passage, Frog and Toad’s desire is conveyed by their continuing to eat the cookies. This characterization aligns with an everyday description of desire. One might say, “How do you know they desired the cookies?,” and receive the response, “Why, because they couldn’t stop eating them!”

Kennett and Smith *precisify* the everyday notion of desire by defining and distinguishing between intrinsic and extrinsic desires and attributing them to Frog and Toad. They *systematize* these precisified notions of desire by stipulating how intrinsic and extrinsic desires operate in the mind of an instrumentally rational agent, together with some additional commitments regarding the relative strengths of different desires. The resulting claim is one that looks like a full-fledged, theoretical account of desire: in an instrumentally rational agent, intrinsic desires are accompanied by the relevant extrinsic desires, where the pairing of intrinsic and extrinsic desires that is overall strongest dictates action.

Kennett and Smith apply this account to *predict* the circumstances under which Frog and Toad would exercise synchronicself-control:

Let’s now ask what Frog and Toad should do. If we interpret this as asking what they *would* do if they were fully instrumentally rational then the answer is clear. Frog and Toad should not eat any more cookies. Their intrinsic desire to be healthy is, after all, stronger than their intrinsic desire to have immediate pleasure, and so the strengths of their extrinsic desires to eat more cookies and not to eat more cookies should simply follow suit (1996, p. 64, added emphasis mine).

Since Frog and Toad are not fully instrumentally rational, Kennett and Smith supplement the view to explain when and how Frog and Toad can *fail* to exercise synchronicself-control.

***3.2 Methodological constraints***

The nature of philosophical folk psychological accounts closely informs the two methodological constraints that govern them. We described the first constraint as having to do with *internal coherence*. The constraint addresses the status of philosophical folk psychology as a theory. While we may in practice tolerate incoherence in lay predictions and explanations of mental states and actions (although we are, in fact, very critical about such cases), full-fledged theoretical accounts are expected to be coherent, i.e., logically consistent. A good philosophical folk psychological theory is thus one where the definitions and relations of mental states are related in such a way as to avoid a contradiction. Kennett and Smith observe this constraint, for example, when they propose the *non-actional* account of synchronic self-control. Kennett and Smith face a potential contradiction between what is supposed to be Frog and Toad’s strongest, temptation-driven desire (to eat cookies) and what is supposed to be their strongest, action-motivating desire (enabling them to exercise self-control). By positing that self-control can be *non-actional*, however, they are able to argue that “there is no contradiction because [Frog and Toad’s] attempt need not itself be an action explained by a desire” (1996, p. 70).

The second constraint has to do with *external* coherence and, more specifically, with *intuitive plausibility*. Since philosophical folk psychological accounts aim to capture what are thought to be real, everyday mental states and real, everyday deliberative and action experiences, the full-fledged theories must cohere with, and make sense of, these everyday mental states and experiences. Kennett and Smith implicitly address the intuitive plausibility constraint (1996, p. 65):

This interpretation of the story is *natural* because failures of instrumental rationality, though perhaps unusual, are not so unusual as to be incomprehensible. *We all know what it is like* when the smell and taste of cookies makes the immediate pleasure of eating them especially vivid and salient, at least when compared with the pallid and inert nature of our knowledge that refraining will lead to good health (added emphasis mine).

Hence, although philosophical folk psychological accounts consist of *precisified* concepts and relations that have been refined to the point of having distinct, technical meanings, they must nonetheless cohere with, and are judged in light of, our everyday views and experiences.

***3.3 Vulnerability to stalemate***

These methodological constraints are not sufficient to help us adjudicate between the three competing views of synchronic self-control, however. There are at least two reasons for this.

First, philosophical folk psychology’s methodological constraints are so minimal that multiple competing theories can satisfy them to a roughly equivalent degree. There are thus no principled means for adjudicating between them. We can return to Kennett and Smith’s analysis of synchronic self-control to illustrate this difficulty. While introducing the two constraints above, I suggested that Kennett and Smith’s proposed, non-actional amendment and the resulting theory are both intuitively plausible and internally consistent. But satisfying the plausibility constraint is not sufficient to recommend either their amendment or their view over either Mele’s or Sripada’s views. The ensuing discussion can go in one of two ways. It can be shown that Kennett and Smith’s amendment and view satisfy both constraints, but then also be shown that, say, Mele’s and Sripada’s satisfy them to an equivalent degree. Alternately, it can be shown that Kennett and Smith’s amendment and view *fail* to satisfy one or both of the constraints, but that so, too, do Mele’s and Sripada’s.

Something like the latter line of reasoning plays out in the literature. Both Mele and Sripada argue that Kennett and Smith’s non-actional amendment and resulting account fail the intuitive plausibility constraint, since non-actional synchronic self-control departs from the everyday conception of self-control as a kind of action. But Sripada then raises a similar objection to *Mele* by arguing that on Mele’s ancillary amendment and resulting account, an agent can only exercise synchronic self-control as a kind of ‘hanger on,’ which does not account for the strongest commonsense form of synchronic self-control which can be initiated independently. As we will see below, a plausibility argument can be raised for Sripada’s own revised view. Neither the intuitive plausibility constraint nor, by analogy, the internal consistency constraint, is thus sufficient to adjudicate between the three leading views of synchronic self-control.

Second, a standard way to adjudicate theories is to assess how well the respective views account for not only our intuitions and everyday experiences, but a broader set of empirical evidence regarding the phenomenon in question. But philosophical folk psychology cannot fully take advantage of this strategy, as its extant constraints are sufficiently lenient to allow philosophers to make claims that are empirically false, or at least claims that we have little reason to accept given current empirical evidence. For example, as will be discussed in Section 5, philosophical folk psychology broadly accepts a *single system* conception of the mind. Single system views of the mind broadly describe a unitary progression of mental states connecting deliberation and action, where mental states proceed serially rather than in parallel (see Davidson, 1970). But our best available computational and empirical theories systematically contradict this assumption of a single system model of the mind. Rather, they support a *multi-system* model of the mind, on which two or more systems operate in parallel to produce action. But neither the intuitive plausibility constraint nor the internal plausibility constraint can account for this feature of our cognitive architecture. Our introspective experiences indicate a unitary progression of mental states rather than a complex system of competing systems, and a single system model of the mind is coherent.

**4. Philosophical folk psychology and its strengths**

We should be careful to note that the problem of adjudication in no way implies that philosophical folk psychology should be eliminated. On the contrary, I want to emphasize that there are very good reasons to subscribe to such an approach to action, although I will introduce an amendment to the view below.

There are numerous metaphysically-oriented reasons for subscribing to philosophical folk psychology. These are widely defended (as are criticisms of these reasons) and emphasize, among other features, philosophical folk psychology’s unsurpassed predictive power. I take these aspects of the issue to be sufficiently well-developed, if continually controversial, to concentrate my efforts here on providing supplemental, *pragmatic* arguments for continuing to subscribe to philosophical folk psychology. These pragmatic arguments are particularly aimed at those philosophers and cognitive scientists who are tempted to reject the role of philosophical folk psychology in explaining deliberation and action.

***4.1 A pragmatic defense of philosophical folk psychology***

The first reason to continue working with philosophical folk psychology stems from general arguments regarding the nature of scientific change. No matter how theoretically dissatisfying one finds philosophical folk psychology, or how theoretically dissatisfying one *expects* to find it in the future, it is reasonable to continue working with the framework while it remains the best and most extensive account of action on offer. One can continue to subscribe to a theory and deem it worthy of investigation even if one no longer expects to accept it in the future. This approach is in line with countless historical instances of scientific change. Despite popular philosophical narratives describing ‘revolutions’ in science, scientific change typically proceeds by means of the gradual, piecemeal revision of individual parts of a theory, including its basic aims, assumptions, and methods. This means that parts of a theory often continue to be held and/or investigated, even as doubts begin to be raised about its other aspects (Laudan et al., 1985).

The second group of reasons for continuing to work with philosophical folk psychology broadly have to do with *utility*. To abruptly set philosophical folk psychology aside in favor of other theoretical approaches would amount to a waste of valuable theoretical resources. It can be argued, for example, that philosophical folk psychology is a valuable starting point for other kinds of theorizing about action, i.e., for computational, psychological, and/or neuroscientific theories. Philosophical folk psychological concepts and relations guide the formulation of novel algorithms, the definition of psychological counterparts, and the search for neural correlates. Along similar lines, it can be argued that many accounts of deliberation and action depend on introspective experiences, and that philosophical folk psychology represents a detailed and systematized repository of such experiences. On this latter line of reasoning, philosophical folk psychology can be thought of as a kind of data set for understanding choice and action, where failing to use it would be to fail to make the most of one’s available research materials.

Notably, both versions of this argument are weaker than the claim that philosophical folk psychology constitutes the *necessary* starting point or data set for other kinds of theorizing about action (Dennett, 1978). This stronger version of the argument can be challenged on the grounds that other kinds of theorizing can reasonably turn to the lay phenomena or lay folk psychological concepts without appealing to the already-theorized philosophical concepts. The pragmatic version of this argument avoids this issue by instead trying to *motivate* the use of philosophical folk psychology, rather than necessitating it.

A third line of reasoning proposes that, irrespective of one’s views on the future of philosophical folk psychology, many of the guiding questions and puzzles pertaining to action are so deeply informed by its assumptions that to ignore them would simply amount to passing these assumptions on uncritically. For example, even if one is interested in the *lay* phenomenon of weakness of will, it is essential to understand how the lay phenomenon has been characterized by philosophical folk psychological accounts before proceeding to other kinds of theorizing. In many respects, weakness of will is no longer just a lay phenomenon. It is now in fact a hybrid, lay and philosophical concept, where it is essential to keep track of its multiple origins and connotations. Hence, this line of reasoning goes, the best way to engage with any inherited assumptions is to make them explicit and, if necessary, to correct them over time.

There are surely other pragmatic reasons for continuing to work with philosophical folk psychology. My aim is not to defend them all here. Most philosophers of action will not need or want to be persuaded of these reasons. Conversely, some cognitive scientists will not be moved by any number of such arguments. By and large, however, there are at least some reasons for thinking that even otherwise-informed approaches to theorizing about action will involve a careful engagement with philosophical folk psychology.

***4.2 Some qualifications***

Still, there are some reasons to qualify our commitment to philosophical folk psychology. First and foremost, it is always reasonable to maintain a principled openness about future refinements or revisions to any given scientific theory. This is true even of the central dogmas in various scientific domains. We may think of philosophical folk psychology as the *truest* theory of action on offer, while remaining open to the possibility that it may need to be refined or revised in the future. We should perhaps even *expect* to refine or revise our concepts of what it means to ‘believe,’ ‘desire,’ and so on.

In addition, it is reasonable to acknowledge that philosophical folk psychology, like any theory, alreadyhas concrete explanatory limitations. We do not need to go so far as to call them explanatory failures (Churchland, 1981). Rather, we can simply point out that there are areas where folk psychology no longer provides the best available explanation of a given issue. Consider the foregoing example of weakness of will. As noted above, philosophical folk psychology plays a central role in formulating and motivating the puzzle of weakness of will. But as has recently been shown, these folk psychological accounts also result in the puzzle’s almost *paradoxical* status ([anonymized]). Once the problem is re-described in a modified form of philosophical folk psychology, weakness of will remains a principled philosophical problem, but one that can be explained and resolved. Not only is weakness of will possible then; on this view, there are *multiple* types of weakness of will. These multiple types make sense of various competing, philosophical explanations of weakness of will, and show that we may require more nuanced ways of addressing the lay phenomenon of weakness of will. There are thus instances where revised versions of philosophical folk psychology offer better explanations of some aspects of or puzzles in action than unrevised versions of philosophical folk psychology do.

*4.3 A naturalistic commitment in philosophical*

On balance, the foregoing arguments suggest that philosophical folk psychology remains our best available theory of action, but nonetheless faces problems of adjudication and, further, can expect to continue to be revised in the future. We can call the resulting view a kind of *incrementalism* about philosophical folk psychology. Importantly, incrementalism recommends that philosophical folk psychology include a meta-theoretical mechanism for adjudicating between both extant and future philosophical folk psychological theories.

I argue that a third methodological constraint should be added to philosophical folk psychology, namely, a *naturalistic constraint*. ‘Naturalism’ is used in many ways in the philosophical literature. The kind of naturalistic constraint I have in mind is much weaker than most kinds of scientific naturalism on offer. It is closest to what Mario de Caro and David Macarthur describe as “respect for the results of the natural sciences” (2004, p. 18). McCauley (2012, p. 187) puts the idea similarly: “all things being equal, philosophical projects in general are pursued more responsibly when they are pursued in the light of the activities, the methods, and the findings of the empirical sciences” (see also Flanagan 2006). What does such ‘respect’ and ‘in the light of’ mean? We can specify three degrees of the naturalistic constraint:

1. The *minimal naturalistic constraint* is satisfied if features of philosophical folk psychology are not directly at odds with our best computational and empirical evidence
2. The *moderate naturalistic constraint* is satisfied if philosophical folk psychology broadly recognizes and incorporate aspects of our best computational and empirical evidence
3. The *comprehensive* naturalistic constraint is satisfied if philosophical folk psychology is constrained by and systematically incorporates our best computational and empirical evidence.

The third, comprehensive constraint suggests that if and when philosophical folk psychology is informed by computational and empirical evidence, it should be informed in a principled and wide-rangingway. The comprehensive constraint implies that philosophical folk psychology should avoid drawing on computational and empirical evidence in a piecemeal manner. Rather, if, for example, our best available evidence indicates that the mind relies on a multi-system architecture, then that architecture should be completely integrated into philosophical folk psychology.

The different versions of the constraint are deliberately modular. The question of which constraint to extend to philosophical folk psychology is precisely what is at stake. Presumably, almost all will agree with the minimal naturalistic constraint. But we should expect far less agreement regarding the moderate and comprehensive versions of the constraint. Which version of the constraint will be favored will depend at least in part on our reactions to those cases in which our intuitions are directly at odds with the scientific evidence. On the one hand, it is reasonable to hold that philosophical folk psychological views must first and foremost account for our intuitions, e.g., the intuition that self-control must be activated and positively sustained by the agent. On the other hand, it is *equally* reasonable to deny that introspection is reliable, i.e., to believe that we do not have accurate self-knowledge of our own attitudes (see Schwitzgebel, 2008 for a review; though see Carruthers, 2011, esp. Chapter 5).

I will not explicitly defend either the moderate or comprehensive constraint here, though I favor the stronger version of the two. Rather, I make a more minimal claim, namely, to suggest that *if* we adopt different versions of these principles, then certain consequences follow, both for philosophical folk psychology in general, and for the puzzle of synchronic self-control in particular.

**5. Applying the naturalistic constraint to extant views of SSC**

In Section 2, we characterized the general strategy for engaging with the puzzle of synchronic self-control as re-describing (1) and thereby showing that (1\*-3) can be made consistent. We can now distinguish between two versions of this more general strategy: to re-describe (1) within a *single*-system model of the mind, and thereby show that (1\*-3) can be made consistent; or to re-describe (1) within *multi*-system model of the mind, and thereby show that (1\*-3) can be made consistent. Single-system views assume a model of the mind on which a single progression of mental states connects motivations and actions, where this progression proceeds serially rather than in parallel. Multi-system models assume two or more semi-autonomous systems operating in parallel rather than serially. In this section, we will see that single-system views are at odds with the minimal naturalistic constraint. By contrast, multi-system views, broadly satisfy the moderate naturalistic constraint but, depending on their specifications, may not satisfy the comprehensive version of the constraint.

*5.1 The naturalistic constraint as applied to single-system views*

The non-actional and ancillary accounts of synchronic self-control are both single-system views. Kennett and Smith present a model of the mind on which a single system provides the basis for a handful of different competing desires (1996). Similarly, recall that Mele identifies three determinants of the motivational force of a desire: a positive motivational base, a negative motivational base, and various cognitive elements that modulate the desire’s motivational strength. But he argues that all three determinants interact within a single progression to connect desires and actions (1987, pp. 68-69).

Single-system accounts face an inherent challenge in explaining synchronic self-control. Since a single progression of mental states is thought to connect motivation and action, some incidental state must be said to disrupt its usual course of development. For example, if an agent most strongly desires to perform some action *A*, but she nevertheless doesn't try to do *A,* then some attitude or event, such as an intervening judgment or intention, must be said to have circumvented or overcome the normally reliable transition.

We find this kind of challenge – and varying solutions to it – in the aforementioned single-system accounts. Kennett and Smith argue that the normally reliable transition between desires and actions can be circumvented by the passive, i.e. *non-actional*, thoughts, modifying desires and actions in such a way as to make synchronic self-control possible. Similarly, Mele suggests that the orthogonal, non-competing desires can be harnessed in the form of ancillary *self-commands* to by-pass the normal transition and thereby exercise synchronic self-control. These views thus overcome the inherent challenge facing single-system accounts, while continuing to satisfy the methodological constraints of internal coherence and external coherence with everyday intuitions.

Nonetheless, single-system accounts face a more difficult challenge when it comes to the naturalistic constraint. Specifically, single-system accounts fail to satisfy even the minimal naturalistic constraint. Recall that the minimal naturalistic constraint requires that features of philosophical folk psychology not *be directly at odds* with our best computational and empirical evidence. But single-system views are precisely at odds with these forms of evidence, which indicate that the mind relies on a multi-system mechanism.

Concretely, there is strong consensus in the decision sciences that at least three computational strategies underwrite three relatively autonomous decisions systems (Dayan and Niv, 2008; Rangel, Camerer, & Montague, 2008; Dayan, 2011; Glimcher and Fehr, 2013). The *hardwired* system is characterized by automatic approach responses to appetitive stimuli and automatic withdrawal responses to aversive stimuli (Macintosh, 1983).[[4]](#footnote-4) Hardwired responses are appropriate in natural environments, since it is broadly beneficial to approach rewards and avoid punishments. However, hardwired responses also lack flexibility, which can result in detrimental decision outcomes (Huys et al.,2012). As a result, this class of responses is typically identified its tendency to persist even in those cases where it is detrimental to do so (Bouton, 2006).

The more sophisticated *deliberative* system explicitly represents possible courses of action and selects those action sequences that promise to maximize value.[[5]](#footnote-5) This procedure is usually illustrated using a ‘decision tree,’ where each node represents a possible choice. Totally value is the sum of the rewards (minus the punishments) along a given branch. The deliberative system searches through the decision tree to find the branch with highest total value.[[6]](#footnote-6) For example, Gary Kasparov may represent three or four upcoming moves in a game of chess, with each possible move further branching into a wide range of subsequent moves. In order to win, Kasparov must represent and choose the best possible sequence of moves overall.[[7]](#footnote-7)

Finally, the third, most computationally complex *habitual* system does not explicitly represent future alternatives, but proceeds by caching positive and negative experiences, assigning values to actions based on their previous total outcomes.[[8]](#footnote-8) Good state-action pairs are those that have produced rewarding outcomes in the past, while bad state-action pairs are those that have produced punishments in the past.[[9]](#footnote-9) To cache experiences, the habitual system employs a feedback signal, which revises the system’s estimates about the environment.[[10]](#footnote-10) For example, a London native caches the best route home from work rather than representing the alternatives every time she commutes.

There is much more to say about these systems. They will be discussed in greater detail in the following sub-section, as well as in Section 6. For the moment, a much more basic claim suffices: the non-actional and ancillary accounts assume a single-system model of the mind. But evidence from our best available comparative psychology, cognitive psychology, and computational and cognitive neuroscience denies this. If we adopt even the most minimal version of the naturalistic constraint, then we should reject both of these explanations of synchronic self-control.

* 1. *The dual-system view fares better*

What of Sripada’s dual-system view? Taking the latter of the two approaches described above, the dual-system view is not directly at odds with multi-system theories of action. It rejects a unitary, serial progression of mental states. It also *recognizes and incorporates* aspects of our best computational and empirical evidence, namely, by introducing multiple, competing systems operating in parallel. It thus satisfies both the minimal and moderate versions of the naturalistic constraint, and so offers a better explanation of synchronic self-control than do either of the two single-system views. Nonetheless, the dual-system view fails to satisfy the *comprehensive* naturalistic constraint.

To see why, let’s take another look at multi-system views. Recall that our best evidence suggests that there at least three, competing sub-systems: an automatic, hardwired system; a representation- or model-based, deliberative system; and a caching, habitual system that carries forward past positive and negative state-action pairs. Because of each of these sub-systems employs a different method, i.e., a different algorithm, for assessing value, each of these systems also provides more and less accurate prediction of value in different contexts. For example, due to its fast, automatic nature, the hardwired system has a high probability of arriving at an accurate prediction in a simple, single-answer problem, such as approaching reward. By contrast, due to its flexible, forward- looking nature, the deliberative system has a high probability of arriving at an accurate prediction in an unfamiliar environment. Thanks to its efficient caching procedure, for its part, the habitual system has a high probability of arriving at an accurate prediction in a complex, familiar context.

Having multiple systems competing to evaluate a single task may seem at best inefficient and at worst disadvantageous. It seems inefficient in the sense that the three systems are potentially redundant, wasting energy computing the same decision problem several times over. It seems plausibly disadvantageous in that the systems may conflict with one another and so hinder the agent’s decision-making overall. But multi-system views are typically equipped with a standard feature that avoids these potential obstacles: namely, they include a principle for arbitrating between these competing systems. Specifically, interactions between systems are thought to be governed by the following ‘accuracy-based principle of arbitration’:

PA Following partial evaluation, that system with the highest accuracy profile, i.e., that system most likely to provide an accurate prediction of expected value, relative to the decision problem at hand, directs the corresponding assessment of value (Deneve and Pouget 2002; Daw, Niv, & Dayan, 2005; Dolan and Dayan, 2013; Lee, Shimojo, & O’Doherty 2014; though see also Keramati, Dezfouli, & Piray (2011) and Pezzulo, Rigoli, & Chersi (2013) for alternative proposals).

According to PA, each of the competing systems partially evaluates the action alternatives. Simultaneously, each system generates an estimate of how accurate its prediction is relative to the decision problem at hand. These estimates, or accuracy profiles, are then compared, and the system with the highest accuracy profile is selected to direct the corresponding valuation task. On the basis of this procedure, then, PA directs valuation based on a system’s accuracy profile, not on its prediction of value. For example, the hardwired system typically coordinates choice in familiar, complex settings, because it typically has a higher accuracy profile in those decision problems, independently of whether it predicts a higher or lower overall value than do either its deliberative or hardwired counterparts.

PA is principled from a computational perspective. It is also extensively supported by both behavioral (see Dickinson and Balleine, 2002 for a review) and neuroscientific evidence in both non-human and human animals (Balleine and Dickinson, 1998; Balleine and Dickinson, 2000; Killcross and Coutureau, 2003; Kiani and Shadlen, 2009; Daw, Gershman, Seymour, Dayan, & Dolan, 2011; Lee et al., 2014). Notably, PA is also consistent with emerging evidence that, for example, the model-free and model-based systems interact (Daw, Gershman, Seymour, Dayan, & Dolan 2011; Dolan and Dayan, 2013; Daw and Dayan, 2014; Otto, Skatova, Madlon-Kay, & Daw, 2014; Gershman, Markman, & Otto, 2014; Pezzulo, Rigoli, & Friston; 2015). That is, proposals investigating interactions and cooperation between the MF and MB systems are consistent with preserving an overall parallel framework, specifically, by suggesting that such information exchanges operate at earlier levels. For example, the model-based system may be used to train the model-free system, but the two systems may nonetheless compete in a moment of decision (Daw and Dayan, 2014; Gershman, Markman, & Otto, 2014).[[11]](#footnote-11) We thus have computational and empirical reasons to believe that multi-system accounts must incorporate some version of PA. But Sripada’s dual-system account does not incorporate PA, or any other principle for arbitrating between systems.

This raises a challenge with respect to the comprehensive version of the naturalistic constraint. Recall that the strongest version of the constraint states that philosophical folk psychology should be constrained by and systematically incorporate our best computational and empirical evidence. In the present context, this means that any philosophical folk psychological view that incorporates a multi-system approach should also incorporate some version of PA. But the dual-system view does not incorporate any version of PA. We must thus conclude that the dual-system fails to satisfy the comprehensive naturalistic constraint.

Lest this seem like petty rule-following, it is important to note that this omission results in considerable, internal challenges for the dual-system view having to do with an agent’s capacity to act in a coherent, unified way despite possessing competing motivational systems. Specifically, the two systems’ independence raises a pair of agency-related difficulties.

The first difficulty pertains to the traditional association between self-control and agency or personhood. Sripada (2010, pp. 795-796) explicitly characterizes the difficulty:

If the divided mind view is correct, then exercises of willpower aren’t truly performed by the agent as a whole, but rather are undertaken by only part of the agent in which only a strict subset of the agent’s full set of desires are active. Common sense understands actions as typically brought about by the agent, not part of the agent, so the existence of ‘sub-personal actions’ of this sort requires a careful philosophical defense and explication.[[12]](#footnote-12)

Sripada deals with this difficulty with relative ease. Almost all actions are the products of some part of our psychology. In any given action, some evaluative commitments and motivations are brought to bear while others remain either irrelevant, or are simply passed over as a consequence of our finite capacities for deliberation and choice. Alternately, an action can be attributed to an agent as a whole when as it *reflects* her overall evaluative commitments.

The second difficulty is less easy to overcome, however. Recall that on the dual-system view, the deliberative and emotional systems issue independent action-desires. But without a principle of arbitration, it is possible, on this view, for an agent to issue simultaneous but contradictory action-desires. For example, Sally can decide to both eat and not eat the key lime pie, and seemingly act in accordance with both of these contradictory action-desires at the same time. This is not a desirable result for a full-fledged theory of action. Problematically, as we will see in Section 6, resolving it will not merely be a question of introducing a principle of arbitration into the dual-system view. Rather, introducing a principle of arbitration will encourage us to revise our understanding of synchronic self-control.

**6. A novel account of action**

In Section 4, I presented several pragmatic reasons for continuing to subscribe to philosophical folk psychology as our best available theory of action. I also argued, however, that, like other full-fledged scientific theories, philosophical folk psychology should remain open to the possibility of future amendments to the view, particularly in the form of refinements or even revisions to the notions of belief, desire, and so on. To this end, I introduced an additional metatheoretical principle, in the form of three versions of the naturalistic constraint, to guide and help adjudicate between various proposed amendments.

In this section, I propose that we are in fact *already* in a position to make such an amendment to philosophical folk psychology and, further, that this amendment has deflationary implications for the puzzle of synchronic self-control. The substance of the amendment substituting replacing our existing, philosophical folk psychological notion of desire with an equivalent concept drawn from the decision sciences. Once we make this amendment, however – satisfying, in the process, the strongest version of the naturalistic constraint – it becomes clear that we should abandon the possibility of synchronic self-control. Instead, we will have to settle for more limited forms of self-control.

*6.1 Turning toward reward*

Philosophical folk psychology broadly subscribes to a *motivational theory of desire*. A motivational theory of desire holds that “to desire that *p* is to tend to be motivated to bring it about that *q*, when one believes that *q* will make it more likely (or will bring it about) that *p*”(Schroeder and Arpaly, 2014, p. 186). This type of theory encompasses the more specific, *dispositional theory of desire*, which holds that "to desire that *p* is to tend to be *disposed* to bring it about that *q,* when one believes that *q* will make it more likely (or will bring it about) that *p* (adapted from Schroeder 2015). These theories have the advantage of being relatively straightforward and of capturing many of the features commonly associated with everyday descriptions of desire. But they face a number of explanatory challenges, including capturing the difference between goodness and desire, explaining why people act out of duty, and so on (Schroeder and Arpaly, 2014).

One prominent solution for dealing with these challenges has been to adopt a reward-based theory of desire (Schroeder, 2004; Schroeder and Arpaly, 2014). Drawing on evidence from reinforcement learning and computational neuroscience, the reward-based theory of desire proposes that, “just as H2O is the unfamiliar essence of water, so… states of the reward system are the unfamiliar essence of desire” (Schroeder and Arpaly 2014, p. 187). Specifically, it holds that “to have an intrinsic desire regarding it being the case that *p* is to constitute *p* as a reward or a punishment” (*Ibid*, p. 187). The view characterizes these concepts of reward and punishment in the senses used in reinforcement learning and associated branches of computational neuroscience (henceforth, the decision sciences). In doing so, the theory preserves the traditional, philosophical folk psychological notion of desire but specifies it in contemporary computational and empirical terms.

My proposed amendment to philosophical folk psychology will hold the same theoretical commitments as the reward-based theory of desire – desire is expressed in terms of the neuroscientific notions reward and punishment – but goes a step further: it proposes that we should *replace* the philosophical folk psychological notion of desire with the technical notions of reward and punishment. The advantage of doing so will be that, rather than nesting the technical notions of reward and punishment within the more familiar philosophical notion of desire, these notions instead play an explicit role in resolving puzzles and debates in the philosophy of action. That is, by substituting the notion of desire with the technical notions of value and reward, we will be able to draw on these notions’ explanatory power directly in order to address our philosophical concerns.

*6.2 Learning, reward, and value*

Standard philosophical approaches describe deliberation and choice in terms of philosophical folk psychology and, by extension, intentional attitudes such as beliefs and desires. The decision sciences recast this area of interest in terms of five core concepts: *goals*, *policies, value, reward,* and *learning*. An agent’s goal is to find an optimal policy that allows her to maximize value, i.e., the sum of the individual rewards she encounters in the different states over time, minus the sum of the individual punishments she encounters in the different states over time. A reward refers to a scalar value representing the degree to which a state or action should be pursued. A punishment refers to a scalar value representing the degree to which a state or action should be avoided. An agent derives such policies from learned experience. In a technical sense, learned experience refers to the forward prediction of various future alternatives and the rewards and punishments that may be associated with each one, together with the ongoing updating of these predictions based on past prediction errors. An agent can use multiple computational algorithms to make these kinds of predictions and updates, where these differing algorithms demarcate relatively autonomous decision systems (Russell and Norvig, 2009).

As sketched in Section 5, our best available empirical evidence suggests that the mind does, in fact, rely on at least three of these computational algorithms and, by extensions, employs three semi-autonomous decisions systems for choice and action. This evidence is both behavioral and neural in nature. Hardwiredresponses are typically identified by their tendency to persist beyond optimality, and they are recognized as distinct behavioral responses in both non-human and human animals. Somatic responses, or automatic bodily reactions such as changes in heart rate and automatic facial expression, are classified as among the most important hardwired responses (Redish, 2013). Hardwired responses are further thought to play a role in psychiatric diseases such as anxiety and depression (Dayan and Huys, 2008; Huys et al., 2011; Huys et al., 2012). Neurally-oriented studies of hardwired responses have focused on three key regions of the brain, namely, the amygdala, the orbitofrontal cortex, and the ventral striatum (Schultz, Tremblay, & Hollerman, 1998; Tremblay and Schultz 1999; Parkinson, Olmstead, Burns, Robbins, & Everitt, 1999; see also Schultz *et al.* 2000, Tremblay & Schultz2000a; Tremblay & Schultz 2000b; Parkinson *et al.* 2002; Paton, Belova, Morrison, & Salzman, 2006; Belova, Paton, & Saltzman, 2008).

Behaviorally, the deliberative and habitual systems, for their part, are typically dissociated using a procedure known as ‘post-training reinforcer devaluation,’ in which a previously desirable reward is devalued – e.g. by overfeeding – and the agent’s response is observed. If the reward is no longer pursued, the agent is thought to be have relied on the deliberative system, thereby recognizing that what had previously been rewarding had devalued. By contrast, if the reward continues to be pursued, the agent is thought to have relied on the habitual system, simply carrying reward’s past value in spite of its recent devaluation. Devaluation has been used in an extensive range of experiments to show the expression of deliberative and habitual responses in both non-human and human animals (Dickinson 1985; Dickinson, Squire, Varga, & Smith, 1998; Dickinson & Balleine, 2002; Tricomi, Balleine, & O’Doherty, 2009). Neurally, the deliberative system is thought to be associated with the dorsomedial striatum, orbitofrontal cortices, and the ventromedial prefrontal cortex (Hampton, Bossaerts, & O’Doherty, 2006; Valentin, Dickinson, & O’Doherty, 2007). By contrast, the habitual system is thought to be associated the nucleus accumbens, the orbitofrontal cortices, the ventral striatum and the orbitoprefrontal cortex (O’Doherty, Dayan, Friston, Critchley, & Dolan, 2003).

Of course, the neural details of these systems do not matter here. They simply support the idea that, whatever our introspective experiences, our mind does, in fact, rely on three competing decision systems. Moreover, to the best of our scientific knowledge, these systems are governed and arbitrated by Daw et al.’s (2005) canonical solution, in the form of the uncertainty-based principle of arbitration, here simply as PA (see Section 5.2). The *implications* of our relying on these three systems will matter, however, particularly for our understanding of the possibility of synchronic self-control. This is because, by making the internal workings of desire more explicit – just as H2O makes the essence of water more explicit – we will see that there is no obvious way to circumvent the puzzle of synchronic self-control.

*6.3 Recasting the problem in terms of value*

Having substituted the philosophical folk psychology notion of desire with the decision-scientific notion of reward – together with its accompanying algorithms and principle of arbitration – we can begin see that there is a general challenge for the possibility of synchronic self-control. Recall the original formulation of the puzzle from Section 1:

1. If an agent *S* most strongly desires to perform some action *A*, and if she believes herself free to *A*, then she will try to *A*, if she does anything at all intentionally.
2. During *I*, *S* most strongly desires to perform *A*, and is free to *A*.
3. Nevertheless, *S* intentionally doesn’t do *A* during *I*.

We can now revise (1) and (2) by substituting it with the amendment to philosophical folk psychology proposed above. On this novel, multi-system account of action, (1) and (2) can be characterized as follows:

1. If S’s *winning decision system D* mostly highly values some option O, and if she believes herself free to O, then S will try do O, if she does anything at all intentionally.
2. During *I*, *S winning decision system D* mostly highly values some option O, and S is free to do O.

To complete the puzzle of synchronic self-control, (3) would also need to be revised. Specifically, it would need to be revised as follows:

1. Nevertheless, *S* intentionally doesn’t do O during *I*.

But 3\* cannot be so revised, because it is precluded by PA.

To see why PA precludes 3\*, recall that each of agent S’s decision systems partially evaluates a set of action alternatives *s* with respect to some decision problem, *p.* Simultaneously, each of S’s decision systems generates an accuracy profile relative to *p*. Each of S’s decision systems’ accuracy profiles relative to *p* are then compared. In keeping with PA, S’s decision system with the highest accuracy profile, *D*, is then selected to complete the corresponding valuation task with respect to *p*. Conversely, the non-winning decision systems are de-activated, i.e., they do not complete the valuation task with respect to *p.* But if this is right, then, once the process of arbitration has taken place and *D* has most highly valued option O, then – barring non-accidental interference – O will be acted on. There is no question of one of the deactivated decision systems *overriding D*, as is permitted on the dual-system view, and thereby preventing S from intentionally doing O during *I*.

Another way of illustrating the tension between PA and synchronic self-control is as follows. Suppose the deliberative system values option O more highly than option Q with respect to *p.* Suppose also that habitual system values option Q more highly than option O with respect to *p.* Finally, suppose that the habitual system reports a higher accuracy with respect to *p* than the deliberative system does, perhaps because *p* amounts to a familiar, i.e., habituated decision problem. Here, a proponent of the *possibility* of synchronic self-control may want to argue, as Sripada does, that since the deliberative system values O most highly, (1) is satisfied, but since S nonetheless does not act on O, then she demonstrates that she is capable of exercising synchronic self-control.

But PA precludes this possibility as well. Again, recall that *D* not only ‘wins’ on the basis of its accuracy profile, while all of the other decision systems ‘lose’ on the basis of theirs. In this example, this means that despite its having the overall highest overall valuation of O prior to arbitration, the deliberative system’s valuation of O necessarily becomes deactivated once arbitration has taken place. But this means that 2\* is no longer satisfied, since O is no longer the most highly valued alternative during *I.*

We thus have good reasons to think that, if we accept the comprehensive naturalistic constraint, and with it the multi-system model of the mind, complete with PA, then we should reject the possibility of synchronic self-control.

**7. Self-control and instrumental rationality**

An agent exercises instrumental rationality to the degree that she adopts appropriate means to achieving her ends. Adopting appropriate means to achieving one’s ends can in turn involve overcoming one’s very strong or even one’s *strongest* desires, i.e., with exercising synchronic self-control. There are thus cases in which the exercise of instrumental rationality requires the possibility of synchronic self-control.

Should we, then, be worried by the foregoing analysis of synchronic self-control? If we adopt the comprehensive naturalistic constraint on philosophical folk psychology and, by extension, the arbitrated, multi-system model of the mind, does this mean we have to give up on a meaningful conception of instrumental rationality?

The conclusion is not so dire, for at least two reasons. First, the name ‘synchronic’ self-control sometimes causes unnecessary confusion with respect to instrumental rationality. There is no doubt that instrumental rationality is closely related to the temporal or ‘in-the-moment’ component of self-control. An agent must be able to act as she sees fit, and to do so *when* she sees fit. In spite its technical title, however, the puzzle of synchronic self-control turns on the superlative nature of a given *desire*, and not on its temporal component. That is, the philosophical puzzle of synchronic self-control turns an agent’s overcoming her *strongest* desire. This *motivational* feature is ruled out by the principle of arbitration dictated by the comprehensive naturalistic constraint. But there is no conflict between the arbitrated, multi-system model of the mind and the *temporal* component of synchronic self-control. This latter kind of self-control is the kind routinely referred to in lay conceptions of control and in the scientific literature, and it remains perfectly possible even once we adopt the comprehensive naturalistic constraint. Hence, although we cannot exercise every possible kind of self-control, we can exercise self-control across a wide-range of circumstances, and we can exercise in-the-moment control of our very strong desires.

Second, it is important to distinguish between ideal instrumental rationality and less ideal versions of the same. Here, we can learn from recent analyses of epistemic rationality. It is widely accepted that we are not, in fact, perfectly rational epistemic agents. Moreover, we do not have to *think* we are perfectly epistemically rational in order to value and cultivate relatively high degrees of epistemic rationality. On the contrary, evidence suggests that we should *not* pretend that we are perfect, since following ideal rules does not always result in optimal outcomes for non-ideal agents (Staffel, 2017).

An analogous approach to instrumental rationality should be developed. As in the epistemic domain, overestimating our capacities for instrumental rationality may well lead us to make assumptions that can, in turn, bring about less optimal outcomes than those we would have achieved if we had accepted our limitations. Conversely, the more we recognize the inner workings of choice and action – and by doing so – the more we recognize the limitations constraining our choices, the more we can plan and choose accordingly, and so act in the most rational ways available to us as real, limited instrumental agents.

**References**

Arpaly, N., & Schroeder, T. (2014). *In praise of desire*. New York: Oxford University Press.

Balleine, B. W., & Dickinson, A. (1998). Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology*, *37*(4-5), 407-419.

Balleine, B. W., & Dickinson, A. (2000). The effect of lesions of the insular cortex on instrumental conditioning: evidence for a role in incentive memory. *Journal of Neuroscience*, *20*(23), 8954-8964.

Balleine, B. W., and O’Doherty, J. P. (2010). Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology* 35, 48–69. doi: 10.1038/npp.2009.131

Belova, M. A., Paton, J. J., & Salzman, C. D. (2008). Moment-to-moment tracking of state value in the amygdala. *The Journal of Neuroscience*, *28*(40), 10023-10030.

Bouton M.E. (2006). *Learning and behavior: A contemporary synthesis*. USA: Sinauer.

Carruthers, P. (2011). *The opacity of mind: An integrative theory of self-knowledge*. Oxford: Oxford University Press.

Churchland, P. M. (1981). Eliminative materialism and propositional attitudes. *The Journal of Philosophy*, *78*(2), 67-90.

Crockett, M. J. (2013). Models of morality. *Trends in cognitive sciences*, 17(8), 363-366.

Coutureau, E., & Killcross, S. (2003). Inactivation of the infralimbic prefrontal cortex reinstates goal-directed responding in overtrained rats. *Behavioural brain research*, *146*(1-2), 167-174.

Davidson, D. (1970). How is weakness of the will possible?. In Davidson, D. (1980), pp. 21-42.

Davidson, D. (1980). *Essays on actions and events.* Oxford: Clarendon Press.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nature Neuroscience, 8(12), 1704-1711.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204-1215.

Dayan, P. (2008). The role of value systems in decision making. In C. Engel & W. Singer (Eds.), *Strüngmann Forum reports. Better than conscious? Decision making, the human mind, and implications for institutions* (p. 51–70). MIT Press.

Dayan, P. (2011). Interactions Between Model-Free and Model-Based Reinforcement Learning,’ *Seminar Series from the Machine Learning Research Group*. University of Sheffield, Sheffield. Lecture recording. <http://ml.dcs.shef.ac.uk/>. Accessed August 2019.

Dayan, P., & Balleine, B. W. (2002). Reward, motivation, and reinforcement learning. *Neuron*, *36*(2), 285-298.

Dayan, P., Niv, Y., Seymour, B., & D Daw, N. (2006). The misbehavior of value and the discipline of the will. *Neural networks,* 19(8), 1153-1160.

Dayan, P., & Niv, Y. (2008). Reinforcement learning: the good, the bad and the ugly. *Current opinion in neurobiology*, 18(2), 185-196.

Dayan, P., & Huys, Q. J. (2008). Serotonin, inhibition, and negative mood. *PLoS Computational Biology*, 4(2), e4.

Dayan, P., & Berridge, K.C. (2014). Model-based and model-free Pavlovian reward learning: Revaluation, revision, and revelation. *Cognitive, Affective, & Behavioral Neuroscience*.

Daw, N. D., & O’Doherty, J. P. (2014). Multiple systems for value learning. In *Neuroeconomics* (pp. 393-410). Academic Press.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204-1215.

Daw, N. D., & Dayan, P. (2014). The algorithmic anatomy of model-based evaluation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1655), 20130478.

De Caro, M., & Macarthur, D. (2004). The Nature of Naturalism. In De Caro, M., & Macarthur, D. (Eds.) *Naturalism in question* (pp. 1-20). Cambridge, MA: Harvard University Press.

Deneve, S., & Pouget A. (2004). Bayesian multisensory integration and cross-modal spatial links. *Journal of Physiology Paris* 98(1-3), 249-58.

Dennett, D. C. (1978). Three kinds of intentional psychology. *Perspectives in the philosophy of language: A concise anthology*, 163-186.

Dickinson, A. (1985). Actions and habits: the development of a behavioural autonomy. *Philosophical Transactions of the Royal Society B: Biological Sciences* 308:67–78.

Dickinson , A., Squire, S., Varga, Z., and Smith, J.W. (1998). Omission learning after instrumental pre-training . *Q. J. Exp. Psychol*. 51, 271-286.

Dickinson, A. & Balleine, B. (2002). The role of learning in the operation of the motivational systems. In Gallistel, C.R. (Ed). *Stevens’ handbook of experimental psychology Vol. 3: Learning, motivation, and emotion*, (pp. 497-533). New York: Wiley.

Dolan, R. J., and Dayan, P. (2013). Goals and habits in the brain. *Neuron* 80, 312–325. doi: 10.1016/j.neuron.2013.09.007

Flanagan, O. (2006). Varieties of naturalism. In Clayton, P. & Simpson, Z. (Eds.) *Oxford handbook of religion and science* (pp. 430-452). Oxford: Oxford University Press.

Gershman, S. J., Markman, A. B., & Otto, A. R. (2014). Retrospective revaluation in sequential decision making: A tale of two systems. *Journal of Experimental Psychology: General*, *143*(1), 182.

Glimcher, P. W., & Fehr, E. (Eds.). (2013). *Neuroeconomics: Decision making and the brain*. New York: Academic Press.

Guitart-Masip, M., Huys, Q. J., Fuentemilla, L., Dayan, P., Duzel, E., & Dolan, R. J. (2012). Go and no-go learning in reward and punishment: interactions between affect and effect. *Neuroimage*, 62(1), 154-166.

Hampton, A. N., Bossaerts, P., & O’Doherty, J. P. (2006) The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *The Journal of Neuroscience*, 26(32), 8360-8367.

Huys, Q. J., Cools, R., Gölzer, M., Friedel, E., Heinz, A., Dolan, R. J., & Dayan, P. (2011). Disentangling the roles of approach, activation and valence in instrumental and pavlovian responding. PLoS computational biology, 7(4).

Huys, Q. J., Eshel, N., O'Nions, E., Sheridan, L., Dayan, P., & Roiser, J. P. (2012). Bonsai trees in your head: how the Pavlovian system sculpts goal-directed choices by pruning decision trees. *PLoS Computational Biology*, 8(3).

Kennett, J., & Smith, M. (1996). Frog and toad lose control. *Analysis*, *56*(2), 63-73.

Kennett, J., & Smith, M. (1997). Synchronic self-control is always non-actional. *Analysis*, *57*(2), 123-131.

Keramati, M., Dezfouli, A., & Piray, P. (2011). Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLoS computational biology*, *7*(5).

Kiani, R., & Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *science*, *324*(5928), 759-764.

Laudan, L., Donovan, A., Laudan, R., Barker, P., Brown, H., Leplin, J., ... & Wykstra, S. (1986). Scientific change: Philosophical models and historical research. *Synthese*, *69*(2), 141-223.

Lee, S. W., Shimojo, S., & O’Doherty, J. P. (2014). Neural computations underlying arbitration between model-based and model-free learning. *Neuron*, *81*(3), 687-699.

Lewis, D. (1972). Psychophysical and theoretical identifications. *Australasian Journal of Philosophy*, *50*(3), 249-258.

Macintosh, N.J. (1983). *Conditioning and associative learning*. Oxford: Oxford University Press.

McCauley, R. (2012). About face: philosophical naturalism, the heuristic identity theory, and recent findings about prosopagnosia. In Gozzano, S. & Hill, C.S. (Eds) *New Perspectives on type identity: The mental and the physical* (pp. 186-206). Cambridge: Cambridge University Press.

Mele, A.R. (1987). *Irrationality: An Essay on Akrasia, Self-Deception, and Self-Control*. New York:

Oxford University Press.

Mele, Alfred. (1992). *Irrationality: An Essay on Akrasia, Self-Deception, and Self-Control*. New York: Oxford Univ Press.

Mele, A. R. (1997). Underestimating self-control: Kennett and Smith on Frog and Toad. *Analysis*, *57*(2), 119-123.

Mele, R. (1998). “Synchronic Self-Control Revisited: Frog and Toad Shape Up,” *Analysis*, vol.

58, no. 4: 305–10.

Mele, A.R. (2003). *Motivation and Agency*. New York: Oxford University Press.

Mele, A. R. (2014). Self-control, motivational strength, and exposure therapy. *Philosophical Studies*, *170*(2), 359-375.

Montague, R. (2006). *Why Choose This Book? How We Make Decisions*. New York: Dutton.

O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., & Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron*, *38*(2), 329-337.

O'Doherty, J. P., Cockburn, J., & Pauli, W. M. (2017). Learning, reward, and decision making. *Annual review of psychology*, *68*, 73-100.

Orlandi, M., & Stroud, S. (*Ms*). Self-Control in Action and Belief.

Otto, A. R., Skatova, A., Madlon-Kay, S., & Daw, N. D. (2014). Cognitive control predicts use of model-based reinforcement learning. *Journal of cognitive neuroscience*, *27*(2), 319-333.

Parkinson, J. A., Olmstead, M. C., Burns, L. H., Robbins, T. W., & Everitt, B. J. (1999). Dissociation in effects of lesions of the nucleus accumbens core and shell on appetitive pavlovian approach behavior and the potentiation of conditioned reinforcement and locomotor activity byd-amphetamine. *The Journal of neuroscience*, *19*(6), 2401-2411.

Parkinson, J. A., Dalley, J. W., Cardinal, R. N., Bamford, A., Fehnert, B., Lachenal, G., ... & Everitt, B. J. (2002). Nucleus accumbens dopamine depletion impairs both acquisition and performance of appetitive Pavlovian approach behaviour: implications for mesoaccumbens dopamine function. *Behavioural brain research*, *137*(1), 149-163.

Paton, J. J., Belova, M. A., Morrison, S. E., & Salzman, C. D. (2006). The primate amygdala represents the positive and negative value of visual stimuli during learning. *Nature*, *439*(7078), 865-870.

Pezzulo, G., Rigoli, F., & Chersi, F. (2013). The mixed instrumental controller: using value of information to combine habitual choice and mental simulation. *Frontiers in psychology*, *4*, 92.

Pezzulo, G., Rigoli, F., & Friston, K. (2015). Active Inference, homeostatic regulation and adaptive behavioural control. *Progress in neurobiology*, *134*, 17-35.

Pool, E. R., Pauli, W. M., Kress, C. S., & O’Doherty, J. P. (2019). Behavioural evidence for parallel outcome-sensitive and outcome-insensitive Pavlovian learning systems in humans. *Nature human behaviour*, *3*(3), 284-296.

Raab, H. A., & Hartley, C. A. (2019). More than two forms of Pavlovian prediction. *Nature* *Human Behaviour*, 3(3), 212-213.

Rangel, A., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature Reviews Neuroscience*, 9(7), 545-556.

Redish, D.A. (2013). The Mind Within the Brain. Oxford: Oxford University Press.

Russell Stuart, J., & Norvig, P. (2009). *Artificial intelligence: A modern approach*. Prentice Hall.

Schroeder, T. (2004). *Three faces of desire*. New York: Oxford University Press.

Schultz, W., Tremblay, L., & Hollerman, J. R. (1998). Reward prediction in primate basal ganglia and frontal cortex. *Neuropharmacology*, *37*(4-5), 421-429.

Schwitzgebel, E. (2008). The unreliability of naive introspection. *Philosophical Review*, *117*(2), 245-273.

Sripada, C.S. (2010). “Philosophical Questions About the Nature of Willpower,”

*Philosophy Compass* vol. 5: 793–805.

Sripada, C. S. (2012). How is Willpower Possible? The Puzzle of Synchronic Self‐Control and the Divided Mind. *Nous*, 48(1), 41–74.

Staffel, J. (2017). Should I pretend I'm perfect?. *Res Philosophica*, *94*(2), 301-324.

Talmi, D., Seymour, B., Dayan, P., & Dolan, R. J. (2008). Human Pavlovian–instrumental transfer. *The Journal of Neuroscience*, 28(2), 360-368.

Tremblay, L., & Schultz, W. (2000a). Reward-related neuronal activity during go-nogo task performance in primate orbitofrontal cortex. *Journal of Neurophysiology*, *83*(4), 1864-1876.

Tremblay, L., & Schultz, W. (2000b). Modifications of reward expectation-related neuronal activity during learning in primate orbitofrontal cortex. *Journal of neurophysiology*, *83*(4), 1877-1885.

Tricomi, E., Balleine, B. W., & O’Doherty, J. P. (2009). A specific role for posterior dorsolateral striatum in human habit learning. *European Journal of Neuroscience*, 29(11), 2225-2232.

Valentin, V.V., Dickinson, A., O’Doherty, J.P. (2007). Determining the neural substrates of goal-directed learning in the human brain. *Journal of Human Neuroscience* 27: 4019- 4026.

1. The author would like to thank Ivan Gonzalez Cabrera, Peter Clutton, Caitrin Donovan, Linus Ta-Lin Haung, Benjamin Henke, Colin Klein, Kathryn Lindeman, Elizabeth Schechter, Julia Staffel, and Ross Pain. Special thanks to Thor Grunbaum. [↑](#footnote-ref-1)
2. At present, the truism captures the principle that an agent need only *believe* that she is free to act in a certain way, as well as unsuccessful attempts at action. Although it is still vulnerable to additional counterexamples, it can, in principle, be revised to accommodate them. My goal is not to establish a version of the truism that precludes all possible counterexamples, however. My goal is to provide a plausible, i.e., empirically-informed principle of motivation, together with an account of synchronic self-control that does not result in a paradox. [↑](#footnote-ref-2)
3. Philosophical folk psychology should be distinguished from other senses of the term ‘folk psychology.’ It is related to, but is not the same as, what can be called ‘lay folk psychology,’ or everyday theories of behavior (regarding the relationship, see Section 3.1). Lay folk psychology is used in everyday explanations of human choice and action, for example, when I ask, “Why was Bob so angry?,” and someone responds, “Because Bob just got a speeding ticket.” This kind of folk psychology takes place in everyday conversations and so on. Similarly, philosophical folk psychology should be distinguished from what are sometimes called ‘mindreading,’ ‘theory-theory,’ and ‘mental simulation,’ respectively. These kinds of folk psychology largely have to do with the mind. Mindreading refers to the cognitive capacities used to explain and predict behavior. Theory-theory is a *theory* about those cognitive capacities. It suggests that the cognitive capacities of mindreading depend on the representation of a kind of theory known as ‘folk psychology.’ Mental simulation is a *different* theory about those cognitive capacities. It suggests that we explain and predict others’ behavior by simulating their mental states. [↑](#footnote-ref-3)
4. In reinforcement learning, this system is formally called the Pavlovian system (for key papers, see Dayan & Balleine, 2002; Dayan, Seymour, & Daw, 2006; Rangel, Camerer, & Montague, 2008; Balleine & O’Doherty, 2010; Huys et al., 2011; Guitart-Masip, Huys, Fuentemilla, Dayan, Duzel, & Dolan, 2012; Huys, Eshel, O’Nions, Sheridan, Dayan, & Roiser, 2012; Daw & O’Doherty, 2014; Gesiarz & Crockett, 2015; see especially Dayan, 2008 and Daw & O’Doherty, 2014). However, the term ‘Pavlovian’ frequently leads confusion among researchers in other fields (for an interesting discussion of how psychologists and machine learning scientists characterize the Pavlovian system differently, see Rescorla’s “Pavlovian conditioning: It's not what you think it is,” 1988). In most fields, as well as in everyday usage, the term ‘Pavlovian’ is usually associated with Pavlov’s original experiments with dogs, where Pavlov trained his dogs by repeatedly ringing a bell and then consistently feeding them afterwards. Famously, the dogs learned to associate the ringing of the bell with the delivery of food, and the dogs’ expectation of food delivery was measured by their salivation. In this context, the food was defined as the unconditioned stimulus, the salivation as the unconditioned response, and the bell as the conditioned stimulus. When the dogs salivated in response to the ringing of the bell, this was identified as the conditioned response. In both public understanding and psychology, then, attention is paid to Pavlov’s discovery of the bell as the conditioned stimulus and its association with the conditioned response, the salivating. By contrast, in reinforcement learning, *it is really the relationship between the unconditioned stimulus (i.e., the food) and the unconditioned response (i.e., the salivating) that is of interest.* One added complication is that Pavlovian *conditioning*, i.e., the learning type, is sometimes also characterized in terms of methods from reinforcement learning methods, i.e., in terms of model-free and model based methods (see especially Dayan and Berridge, 2014; Pool, Paul, Kress, O’Doherty, 2019; see also Raab and Hartley, 2019). Thanks to an anonymous reviewer for pressing me on this last point. [↑](#footnote-ref-4)
5. The deliberative system is formally called the goal-directed or model-based system (e.g. see Dayan, 2011). [↑](#footnote-ref-5)
6. Hence its alternative name, tree search. [↑](#footnote-ref-6)
7. Example from Peter Dayan (2011). [↑](#footnote-ref-7)
8. The habitual system is formally called the habit-based or model-free system (e.g. see Dayan, 2011). [↑](#footnote-ref-8)
9. This formulation is owed to Crockett (2013). [↑](#footnote-ref-9)
10. The feedback signal works much like exclamations of ‘Hotter!’ and ‘Colder’ in the children’s game Hot-or-Cold. The Seeker moves around the room with the general goal of finding a hidden object. The Hider helps the Seeker by telling her whether she is getting closer or farther away. The Hider’s suggestions operate like an error signal by helping the Seeker refine her predictions, albeit without giving her detailed instructions about where to go (analogy from Montague, 2006). [↑](#footnote-ref-10)
11. For instance, in addressing precisely this question, Daw, Gershman, Seymour, Dayan, & Dolan (2011, p. 1211) observe that that such a hybrid approach of interaction-before-arbitration “would appear to remain consistent with the lesion data suggesting that the systems can function in isolation (Killcross and Coutureau, 2003; Yin et al., 2004, 2005), and with behavioral data demonstrating that distinct decision systems may have different properties and can be differentially engaged in different circumstances (Doeller and Burgess, 2008; Frank et al., 2007; Fu and Anderson, 2008). It also remains consistent with other fMRI studies (Doeller et al., 2008; Poldrack et al., 2001; Venkatraman et al., 2009) suggesting that overall activity in different brain systems associated with either system can modulate with time or circumstances, presumably in relation to the extent that either process is engaged.” [↑](#footnote-ref-11)
12. Elsewhere, Sripada characterizes the same objection in terms of personhood (2014, p. 22). [↑](#footnote-ref-12)