Garson, J. (2016) "Two Types of Psychological Hedonism," *Studies in History and Philosophy of Biological and Biomedical Sciences* 56: 7-14.

Title: Two Types of Psychological Hedonism

Abstract: I develop a distinction between two types of psychological hedonism. *Inferential* hedonism (or "I-hedonism") holds that each person only has ultimate desires regarding his or her own hedonic states (pleasure and pain). *Reinforcement* hedonism (or "R–hedonism") holds that each person's ultimate desires, whatever their contents are, are differentially reinforced in that person's cognitive system only by virtue of their association with hedonic states. I'll argue that accepting R-hedonism and rejecting I-hedonism provides a conciliatory position on the traditional altruism debate, and that it coheres well with the neuroscientist Anthony Dickinson's theory about the evolutionary function of hedonic states, the "hedonic interface theory." Finally, I'll defend R-hedonism from potential objections.

Section 1. Introduction.

The English philosopher and political theorist Jeremy Bentham wrote, in 1780, "Nature has placed mankind under the governance of two sovereign masters, pain and pleasure." (Bentham 1789 [1780]). Since that time, many philosophers have taken Bentham's dictum as a classic statement of the view called psychological hedonism (in contrast to ethical hedonism) (see Feinberg 1987, 1; Sober and Wilson 1998, 1; Stich et al. 2010, 152 [*fn.* 10]). Yet this dictum contains a fundamental ambiguity, one that has not yet been recognized. Let *A* be an agent, and *D* some desire that *A* has – a desire other than the desire that *A* obtains pleasure or avoids pain.[1] In order for *A* to have *D*, must *A believe* that satisfying *D* will contribute to pleasure? Or is it enough that the satisfaction, or even the mere existence, of *D* is, in fact, pleasurable, and this fact causes the desire to persist?

I will call the first kind of hedonism "inferential hedonism," for reasons to be explained in the next section. (Alternately, I will just refer to it as "I–hedonism.") I–hedonism holds that for any agent, *A*, and for any desire, *D*, *A* has *D* only because *A* believes that the satisfaction of *D* will promote *A*'s pleasure. In this view, in order for *A* to desire something other than pleasure, then, *A* must possess certain beliefs about the relationship between the satisfaction of that desire and pleasure. In most cases, these will be causal beliefs (i.e., that the satisfaction of *D* will cause pleasure). They can also be "constitutive" beliefs, that is, beliefs to the effect that satisfying *D* is constitutive of pleasure (e.g., my belief that health is somehow constitutive of happiness). This is the kind of hedonism that philosophers are typically thinking about when they discuss psychological hedonism.

I will call the second kind of hedonism, "reinforcement hedonism" (or, alternately, "R-hedonism"). R-hedonism holds that, where *D* is an ultimate desire, *D* is maintained or reinforced in *A*'s cognitive system only by virtue of the fact that *D* is associated with pleasure. When I say that *D* must be "associated with" pleasure, I am thinking of two different sorts of cases. In the first case, the satisfaction of *D* (regularly, typically, or non-

[1] In the following, I'll use the term "pleasure" as an abbreviation for, "pleasure, or the avoidance of pain."

negligibly) causes, or is constitutive of, pleasure. In the second case, *A* derives pleasure merely from *entertaining* the satisfaction of *D*. According to R–hedonism, it is possible for someone to have a long-standing, ultimate desire that is never satisfied, such as a desire for revenge or a desire for world peace. The R-hedonist simply maintains that such desires are reinforced because the agent derives pleasure from imagining their being satisfied. A monk can have a lifelong, unfulfilled, and ultimate desire for sex. The R-hedonist says that the only reason this desire is reinforced is because the monk derives pleasure from contemplating its satisfaction. When I contemplate satisfying a desire, and I get pleasure from that, that sets up a kind of "virtual reinforcement scheme" that causes the desire to persist. (Note that the R-hedonist is not committed to the claim that all desires are reinforced only by virtue of their association with pleasure, but only that "ultimate" desires are reinforced this way. "Instrumental" desires are maintained simply by virtue of the agent's beliefs about the relation between the instrumental and ultimate desire.)

Another way of framing the distinction between I–hedonism and R–hedonism is in terms of the distinction between the *content* of a desire, on the one hand, and the *mechanism* by which that desire is reinforced in the cognitive life of the agent, on the other (or, alternatively, the *function* of that desire – see below). I–hedonism is a theory about the contents of one's ultimate desires. It claims that one only has ultimate desires about one's own hedonic states. R–hedonism is a theory about the mechanism by which those desires are maintained or reinforced over time – namely, by virtue of their actually being associated in the right sort of way with one's hedonic states. According to R-hedonism, people can have ultimate desires regarding the welfare of others. R-hedonism just holds that, if those desires were not, in fact, associated with pleasure, they would soon disappear. LaFollette (1988) suggested a similar distinction, though he did not consider the latter view to be a variety of hedonism and he did not focus narrowly on pleasure, per se, as the sole reinforcement mechanism, but rather what he called "satisfaction."

One purpose of the following is to clarify the distinction between the two types of hedonism, and to situate the distinction in relation to the traditional altruism-egoism debate. It is not merely, however, an exercise in conceptual clarification. A second goal is to provide some biologically and psychologically plausible reasons for rejecting I–hedonism and accepting R–hedonism.

The following consists of six sections. After the introduction (Section 1), I will clarify the distinction between I-hedonism and R-hedonism, particularly with respect to the traditional egoism-altruism debate (Section 2). In Section 3, I'll review Sober and Wilson's (1998) evolutionary argument against I-hedonism and explain why I find it convincing. In Section 4, I'll provide an empirically-oriented argument for R-hedonism, namely, that it receives support from the neuroscientist Anthony Dickinson's theory about the biological function of pleasure. In Section 5, I will defend R–hedonism against a host of potential objections. In the final section, I'll make some concluding remarks and gesture toward some further lines of inquiry.

Section 2. I-hedonism and R-hedonism.

The distinction between I–hedonism and R–hedonism is best understood in the context of the traditional altruism–egoism debate. The traditional altruism debate emerges at the intersection between two distinctions: that between “ultimate” and “instrumental” desires, and that between *self*–directed and *other*–directed desires. To understand the altruism debate, in its traditional form, one must understand these two distinctions (see Garson 2015, Chapter 1 for an overview).

Let *A* be an agent, and *D* be some desire that *A* has. *D* is an *instrumental* desire if and only if the only reason *A* has *D* is that *A* believes the satisfaction of *D* will promote the satisfaction of some other desire, *D′*. (When I say one desire “promotes the satisfaction” of another, I mean either that the satisfaction of the first causes the satisfaction of the second, or that the satisfaction of the first is somehow constitutive of the satisfaction of the second.) *D* is an ultimate desire if and only if it is not instrumental. Another way of approaching the distinction is by imagining that an agent’s desires form a ladder–like hierarchy. *A*’s “ultimate” desires are simply those at the top of that hierarchy. Ultimate desires would continue to exist even if the agent did not believe that their satisfaction would promote the satisfaction of others. Note that an agent can have more than one ultimate desire. It is also possible that an agent’s ultimate desires conflict with each other. Finally, an agent’s ultimate desires can change over time; a desire can “convert” from being instrumental to being ultimate, or vice versa.

Note that I do not have a special theory here about what constitutes a belief or a desire. For example, must desires be something like propositional representations? Must the agent’s beliefs have the right sorts of formal or syntactic structure in order to constitute genuine beliefs? That would exclude most non-humans, and even some human beings, from having “beliefs.” Or, can these beliefs be more rudimentary belief-like states, for example, along the lines of what Kim Sterelny calls “decoupled representations” (Sterelny 2003, Chapter 3)? Clearly, slightly different versions of I-hedonism can be generated depending on how one explicates the notions of belief and desire.

The distinction between self–directed and other–directed desires is a distinction regarding the *contents* of a person’s desires, that is, what they are about. *D* is other-directed for *A* if it is about the welfare of some other agent, *A′*. *D* is self-directed for *A* if the desire is about *A*’s welfare. Note that a desire can be both self- and other-directed, such as my desire that *my wife and I buy a house*. Moreover, a desire can be neither self- nor other-directed, such as a desire that *the universe persist forever*. (Of course, people might disagree about what constitutes “welfare,” and I have no special theory here. I hope that the examples serve to illustrate, at least crudely, the distinction I am trying to capture.)

Putting these two distinctions together, one can formulate the traditional altruism–egoism debate. The traditional egoist holds that *all ultimate desires are self-directed*. For example, the traditional egoist maintains that people only have ultimate desires for things like happiness, health, wealth, respect, or power. This position does not imply that people never have other–directed desires. The egoist simply holds that, to the extent that they do, those desires are instrumental and not ultimate. The traditional altruist holds that, perhaps in addition to ultimate, self–directed desires, people sometimes have ultimate, other–directed desires.

Traditionally, hedonism is construed as a special variety of egoism (this is the variety of hedonism that I will refer to as "I-hedonism" for reasons to be explained in the next section). For the traditional hedonist, all ultimate desires boil down to the desire for pleasure. People clearly do have desires for things like wealth, health, or power, but only because they believe that those things will contribute to pleasure. Hedonism, in this sense, strikes me as the most plausible form of egoism. That is because it is hard to see why somebody would want things like power, wealth, and so on, unless that person believed that having those things would feel good, or be pleasurable. Of course, the question of what exactly "happiness," or "pleasure," or "feeling good," amounts to, is an empirical question that will be progressively illuminated by psychology and neuroscience (see Kringelbach and Berridge 2010 for a good starting point on the neuroscience of pleasure). It strikes me as unfair to demand that the hedonist provide a perfectly lucid account of what, precisely, "happiness" or "pleasure" amounts to, prior to the development of the relevant empirical research program.

Having set up the structure of the traditional altruism debate, one can now distinguish easily between two types of hedonism. "Inferential" hedonism, or I–hedonism, is just traditional hedonism. It is the view that people only have ultimate desires regarding their own pleasure. I call it "inferential" hedonism because it emphasizes the inferential role that ultimate desires play in generating new (instrumental) desires. Ultimate desires, in this view, interact with beliefs, typically causal beliefs, in an inference–like manner. In doing so, they generate a multiplicity of instrumental desires. Another way of thinking about it is that, according to I–hedonism, when an agent desires something other than pleasure, for example, a warm meal, or a soft bed, or a friend's recovery, it is because the agent has made an inference regarding the likely consequences of satisfying that desire. In the worldview of I-hedonism, humans are calculating machines. The crucial point is that for I-hedonism, when one forms a desire for the welfare of another person, it is because the person has made a kind of inference, and this inference explains why the desire exists.

Reinforcement hedonism, or R–hedonism, is neutral about the contents of one's ultimate desires. Unlike I–hedonism, it is not a theory about what our ultimate desires are about. Rather, it states that one's ultimate desires, whatever they are about, are reinforced in one's cognitive system only by virtue of being associated with pleasure – regardless of the agent's beliefs about that association. Reinforcement hedonism is perfectly consistent with the possibility that people have ultimate desires regarding the welfare of others. It just holds that, if those desires were not, in fact, associated with pleasure, the agent would soon stop having them. According to R-hedonism, other-regarding desires need not persist by virtue of anything like an *inference* that connects the representation of the desire as being satisfied, and the representation of the agent's own happiness. The view that I call "R-hedonism" was prompted by Hugh LaFollette's (1988, 504) remark that, though egoism is false, it contains the germ of truth that, "we are psychologically constituted to decrease an activity, including moral activity, if we are not ultimately satisfied by it."

For some, to say that a certain feature of one's cognitive life was reinforced by virtue of its association with such-and-such, implies that that item has the *function* of bringing about such-and-such (e.g., Millikan 1984, 27). If one accepts this construal of "function," then one can neatly distinguish the two types of hedonism in terms of the distinction between the content, and the function, of a desire. To make an analogy to natural selection, one can imagine that, at any given time, an agent has a "population" of ultimate desires. Some of those desires can be innate, and some of those desires can be acquired (assuming that the innate-acquired distinction turns out to be a sensible one to make). Within that initial "pool," some desires, by virtue of being associated with pleasure, tend to be reinforced, or maintained, in the agent's cognitive life. Other desires, because they are not associated with pleasure, are eliminated. The analogy to natural selection, of course, is far from perfect, partly because desires do not really "reproduce." I am not concerned here with making the analogy perfect (see Garson 2011; 2012; 2015 for discussion of function and selection processes).

Finally, there are two distinct ways in which a desire can be "associated with pleasure." First, a desire can be "associated with pleasure" because its satisfaction causes, or is constitutive of, pleasure. If this were the only way in which a desire could be "associated with pleasure," however, then R–hedonism would not be a very convincing view. After all, it seems intuitively plausible that people can maintain certain ultimate desires for indefinitely long periods of time, even if those desires are never satisfied. One can have a lifelong, and perhaps even an ultimate, desire for world peace, even if that desire is never satisfied. A monk can have a lifelong, and perhaps even an ultimate, desire for sex, even if that desire is never satisfied. Fortunately for R-hedonism, there is a second way in which a desire can be associated with pleasure. Sometimes, a desire can be associated with pleasure simply because the agent derives pleasure merely from *representing that desire as being satisfied*. For example, I can form the desire to see my children after a long period of absence, and I can derive pleasure merely by *imagining* seeing them, that is, merely by representing the satisfaction of that desire. According to R-hedonism, this constitutes one way for a desire to be "associated with pleasure." Note that R-hedonism is based on highly contingent empirical assumptions. I consider this a strength of the theory and not a weakness.[2]

[2] Here is one way that R-hedonism could potentially be falsified. If it were discovered that *both* pleasure *and* the reinforcement of desire had a common cause, and one is not causally implicated in the other, that would seem to undermine R-hedonism. That is because, in that scenario, pleasure is not the cause of desire reinforcement but the two are results of the same cause. For example, some people believe that dopamine signals indirectly cause *both* pleasure, *and* reinforcement learning, but that the pleasure is not the cause of the reinforcement. I do not believe that the current evidence we have is decisive. In particular, there are still open questions about the precise role of dopamine in reinforcement learning (see Berridge and O'Doherty 2013 for conflicting views about the role of dopamine).

R-hedonism is even logically consistent with the possibility that a person has an ultimate desire for something, *while believing* that the satisfaction of that desire will not be pleasurable. Again, it only requires that the agent derive pleasure from representing the satisfaction of the desire. A desire for revenge, for example, can have this paradoxical quality. One can have a strong desire to take revenge on another, and even take pleasure in imagining it, while knowing that doing so would probably not be enjoyable. The abuse of L-DOPA medication in some Parkinson's patients might also exemplify this possibility, as some patients seem compulsively to want more medication despite the fact that they do not greatly enjoy it and despite the fact that they might recognize this (Berridge 2007, 397). This data is at the very least problematic for I-hedonism, but it is consistent with R-hedonism.[3]

Section 3. Two Arguments Against I-hedonism.

There are at least two main problems for I-hedonism, from an empirical perspective. These problems do not defeat the view entirely, but they strongly suggest that I-hedonism is not empirically well-motivated. Sober and Wilson (1998) complain, rightly, that in many discussions, I-hedonism (or more generally, egoism) is treated as a sort of "default" position – the "null" hypothesis, if you will. The burden of proof is placed on the altruist to demonstrate the existence of genuine, other-directed desires, rather than on the egoist to demonstrate their non-existence. But there is no compelling biological or psychological motivation for treating I-hedonism as a kind of "default" hypothesis. Neither side of the debate should be seen as uniquely having a burden of argument.

[3] This is perhaps what Schroeder et al. (2012, 106) mean when they say that, "one of the more interesting implications suggested by the neuroscientific work is that hedonism is false…A person can have an intrinsic desire to act in a certain way even in the absence of any pleasure (or pain) signal." I do not think the neuroscientific evidence is quite as damaging for I-hedonism as Schroeder et al. (2012) suggest. For example, despite the evidence from Parkinson's patients, there do not appear to be any well-documented examples in the neuroscientific literature of a person who has a strong desire for something but who also believes that they will derive *no pleasure whatsoever* from satisfying it (Berridge, pers. comm). An important upshot of Kent Berridge's work on motivation is that, in human beings, "wanting" and "liking" are somewhat independently manipulable and they seem to have different neurobiological mechanisms (mesolimbic dopamine tract in the case of wanting and various opioid and endocannibinoid receptors in the case of liking – e.g. Berridge 2010). This would suggest that it is at least theoretically possible, in a human being, to dissociate these entirely. Similar such dissociations have been carried out in rats (Robinson and Berridge 2013). But this does not imply that, in human beings, having a desire for something and expecting to derive enjoyment from its satisfaction can ever, *in fact*, come apart. It seems to me that in order to decisively refute I-hedonism, one would have to document, in a human, a complete dissociation between wanting and anticipated pleasure.

The first, and most well-known, of these problems is the "experience machine" thought experiment devised by Robert Nozick (see Nozick 1974 for discussion). But I am going to turn my attention to a more empirically-oriented, evolutionary argument against I-hedonism. Sober and Wilson (1998) (also see Sober 1994) develop the first problem lucidly, particularly in the final chapter of *Unto Others*. Their argument, in essence, is very simple. They argue, roughly, that altruists would make for better parents than hedonists, and that altruism is no more energetically costly to maintain. As a consequence, altruism is more likely to have evolved than hedonism as a psychological mechanism for child-rearing.[4] I will develop the argument more cautiously here, since I will return to it in Section 5. I realize my rendering of their argument below is very curtailed. See Schulz 2011 and Garson 2015, Chapter 2 for a fuller overview and discussion.

Their argument stems from two considerations. First, parents typically care about their children. Second, natural selection has brought it about that parents typically care about their children. This raises a question: what psychological mechanism might natural selection have used to motivate parents to care for their children? There are at least two plausible such mechanisms. The first is the "altruistic" solution: give parents ultimate desires regarding the welfare of their children. The second is the "hedonistic" solution: give parents ultimate desires regarding their own hedonic states, but "wire" them in such a way that they tend to derive pleasure from caring for their children. (Of course, natural selection might have used some combination of these mechanisms, but that would still constitute a form of altruism, so I will set that aside.) So, which of the two solutions is more likely to have evolved by natural selection?

Sober and Wilson argue that when evaluating the relative likelihood that natural selection would promote an altruistic motivational architecture, rather than a hedonistic one, we must consider three factors: availability, energetic efficiency, and reliability (Sober and Wilson 1998, 305; also see Sober 1994). First, *availability*: do we have any reason to think that altruism was, in fact, available to ancestral populations, for example, that the right sort of genetic mutations would have occurred that underpin it? Sober and Wilson think that altruism would have been just as readily available as hedonism to ancestral populations. This is because both altruism, and hedonism, utilize the same underlying psychological architecture – namely, a cognitive system that involves beliefs and desires.

The second consideration is *energetic efficiency* (or what Sober 1994 calls "side constraints"). Crudely put, do we have any reason to think that altruism is more metabolically costly to maintain than hedonism? Does altruism require the consumption of more calories? They argue that with respect to energetic efficiency, the two are

[4] Note, importantly, that Sober and Wilson do not merely intend to argue for the existence of parental altruism. Rather, they intend to make a much more general point about the evolution of psychological altruism by group selection, where that altruism is directed toward other members of the group and not just offspring (Sober and Wilson 1998, 326). However, for my purposes I am going to restrict attention to the more limited version of their argument, which focuses on the relation between parents and children.

probably comparable. That is because they merely represent differences in the contents of our ultimate desires, rather than the underlying mechanisms that sustain our desires (Sober and Wilson 1998, 322). This is clear if one thinks that desires and beliefs involve something like an inner representation of a proposition. There is no reason to think that my ability to represent the proposition, *you eat ice cream*, requires more exertion than my ability to represent the proposition, *I eat ice cream*.

Third, they argue that altruism is more *reliable* than hedonism in making sure parents take care of their children. This is because the hedonistic motivational structure is much more complicated than the altruistic motivational structure. Since it has more “moving parts,” it is more likely to break down. Another way of putting it is that, unlike the altruistic solution, the hedonistic solution makes the desire to care for one’s offspring *contingent* on its effect (or perceived effect) on one’s own hedonic states. As a consequence, it introduces a source of fragility that altruism does not possess. In short, the altruist solution is superior to the hedonist solution with respect to reliability, and it is comparable with respect to availability and energetic efficiency.[5] Hence, it is more likely to have evolved by natural selection.

Stich (2007, 276) points out that Sober and Wilson’s argument does not *guarantee* that altruism would have evolved by natural selection. After all, only the most die-hard adaptationist would infer that, just because one solution is in some sense more “optimal” than another, natural selection would have ensured its appearance. I agree with Stich’s point, but I don’t think it is devastating for Sober and Wilson’s argument. Sober and Wilson’s argument simply intends to provide us with one line of empirical evidence for altruism; it is not meant to be decisive. The point is that we can use Sober and Wilson’s argument, in conjunction with other lines of evidence (drawn from, say, neuroscience or social psychology) to provide a convincing case for altruism (see Schulz 2011).

More importantly, Stich invokes the existence of sub-doxastic states to argue that altruism and I-hedonism might, in fact, be comparable with respect to reliability (280; also see Stich et al. 2010, 168). The idea is this: suppose that I-hedonism is true. Then, the only reason I care about the welfare of my children is because I believe that *I will derive pleasure from promoting their welfare.* Suppose, however, that this belief has a “sub-doxastic” status, that is, it is so deeply entrenched in my cognitive system that countervailing evidence cannot dislodge it (see Stich 1978). Then, Stich claims, I would be just as reliable a caretaker as the altruist parent. In short, depending on the details of the corresponding architectures, we might expect the I-hedonist and the altruist to behave in exactly the same way with respect to rearing children. To the extent that they do, there would be no fitness differences between them.

I suspect that Sober and Wilson can avoid the force of this critique by pointing out that, even if Stich’s considerations *minimize* the reliability gap, they don’t close it entirely.

[5] Another way of putting the point, in the language of multiple-criterion decision-making, is that the altruistic solution weakly “dominates” the hedonistic solution (see, e.g., Sarkar and Garson 2004).

The point is that the motivational architecture that Stich describes is more complex than the motivational architecture that Sober and Wilson describe. (In Stich's hypothesis, parent *P* has an ultimate desire about *P*'s own welfare, a subdoxastic belief about the relation between *P*'s pleasure and the welfare of the offspring, and an instrumental desire for the offspring's welfare. In Sober and Wilson's hypothesis, the agent merely has an ultimate desire about the welfare of the offspring. We are supposing that the two agents are pretty much alike in all other respects, for example, their empirical knowledge about what is beneficial for the offspring.) Strictly speaking, there are simply more ways that the hedonist's motivational architecture can fail to perform its stipulated function of child rearing (for example, by a genetic mutation that prevents the proper subdoxastic belief from developing). Hence, it would be at least *slightly* less reliable. One could try to argue that the fitness differences between the hedonist and altruist architectures are so small as to be negligible. But even very small fitness differences can have major evolutionary consequences over long time-scales.

In short, I find Sober and Wilson's argument to provide a strong, though defeasible, piece of evidence for altruism. At any rate, I am going to set this argument to one side, because in Section 5, I will argue that even if one accepts Sober and Wilson's argument, it cannot be converted in any straightforward way into an argument against R-hedonism, which is what I am really concerned to defend.

Section 4. Two Arguments For R-hedonism.

In this section I will offer two arguments for R-hedonism. The first is more conceptual and the second is more empirical. The first is that R-hedonism provides a conciliatory perspective on the traditional altruism debate. The second is that it coheres well, and is even supported by, our evolving grasp of the neuroscience of pleasure. I'll take each of these in turn. These arguments are not meant to be compelling, but to provide some preliminary support and get the theory "off the ground." Since R-hedonism is an empirical conjecture, one must ultimately accept or reject it on the basis of empirical evidence.

The traditional altruism debate has, for centuries, been sustained by two apparently contrary sets of intuitions about human motivation. One strength of R-hedonism is that it is consistent with both sets of intuitions. The first, articulated by Joseph Butler in his *Fifteen Sermons Preached at the Rolls Chapel* of 1726, suggests that I-hedonism misconstrues the role of pleasure in human motivation. It seemed obvious to Butler that pleasure is typically a by-product or outcome of the satisfaction of a desire for something else – something other than pleasure – rather than being part of the very content of the desire. As he put it, "That all particular appetites and passions are towards external things themselves, distinct from the pleasure arising from them, is manifested from hence; that there could not be this pleasure, were it not for that prior suitableness between the object and the passion..." (1729, 204; see Sober 1992 for discussion). In other words, Butler agreed that the satisfaction of a desire typically has a positive impact on our hedonic states. Yet he thought we could make sense of this intimate connection between desire satisfaction and hedonic states without assuming that our ultimate desires are only *about* those hedonic states. If one accepts R-hedonism, one could agree entirely with Butler's view that I-hedonism simply misconstrues the motivational role of pleasure.

In apparent contrast to Butler, Bentham wrote, "Nature has placed mankind under the governance of two sovereign masters, pain and pleasure." As I noted above, commentators have interpreted this as an endorsement of I-hedonism, but it can be interpreted differently. One could interpret this claim to mean that ultimate desires, whatever their contents are, are reinforced only by virtue of their association with pleasure. Intuitively, this would still constitute a way of being under the "governance" of pleasure and pain. The point here is that if we accept R-hedonism, we would have to endorse Bentham's intuition that pleasure and pain play a regulatory role in all human motivation. There would be no "additional" realm of human motivation, no matter how refined or noble, that would be exempt from this law. The fact that R-hedonism can accommodate both "Butlerian" and "Benthamite" intuitions about motivation strikes me as a strength of the view. I'll return to this point in the final section.

The second argument is more empirical. R-hedonism is supported by our evolving understanding of the neuroscience of pleasure. One of the central theories in this area is Anthony Dickinson's view, which he calls the "Hedonic Interface Theory" (HIT) (see Dickinson and Balleine 2010; Dickinson 2008). According to Dickinson, human behavior is governed by two separate "psychologies," the intentional (belief-desire) psychology, and the stimulus-response psychology (or what he calls the "beast machine"). The intentional psychology consists in the representations of various goals, and the representations of causal relations between them. Simply put, it is made up of beliefs and desires, or at least, states that are belief-like and desire-like. The stimulus-response psychology is built of fixed action patterns, conditioned associations, and instrumental habits.

A crucial feature of Dickinson's dual-psychology view is that these two psychologies can, as it were, run counter to one another in the way they influence the creature's behavior. Specifically, though the "stimulus-response" psychology is typically grounded in the needs of the organism, the intentional psychology can rapidly generate goals that are not so grounded. We can invent biologically destructive goals; we do so all the time. So, if we have two psychologies, we need a system in place that can somehow reconcile the two, and specifically, ensure that the intentional psychology serves the needs of the organism.

In his view, this is where pleasure and pain come in. Pleasure is not typically part of the *content* of what we want. It is part of a mechanism that causes certain "wants" to be reinforced, and others eliminated, depending on how they serve the needs of the organism. For Dickinson, the biological function of pleasure and pain is to *modulate the value we place on represented goals, in such a way as to align the intentional psychology and the stimulus-response psychology*. The function of pleasure and pain is to harness the intentional psychology and force it to serve the biological goals of the organism. They do so by serving as a reinforcement mechanism for various goals. Dickinson puts the point in the following way:

> The function for conscious hedonic and affective experience is to act as a motivational interface between the psychologies of the cognitive creature and the reflex machine. The function of this interface is to ground intentional desires, or in other words, cognitive representations of goal values in the biological responses of the reflex machine to motivationally relevant variables, such as nutritional and fluid depletion, poisoning, hormonal states, body temperature, and so on. The motivational imperative of this experience then changes the representation of goal value in the cognitive psychology so that subsequent goal-directed actions are controlled by this new value (Dickinson and Balleine 2010, 75).

One of his important pieces of experimental work, the "Palermo protocol," shows how, on the level of intentional psychology, an organism can have a desire for *X* (can represent *X* as a goal), while on the level of the stimulus-response psychology, it can also have an aversion toward *X*. One way to set up an aversion to *X* is to allow, say, a rat, to experience *X*, and then, a few hours later, induce nausea in the rat. This would represent a kind of "conflict" between the two psychologies, as the rat now has a desire for something that will make it feel deathly ill. For example, in one version of this scenario, he allowed a rat to drink sugar water, and then, hours later, induced nausea in the rat. This created an association, at the level of the "stimulus-response psychology," between sugar water and nausea, such that the next time the rat consumed sugar water, it would feel nauseous immediately. But the rat did not "know" that sugar water would immediately induce nausea; it still represented the sugar water as desirable and pursued it. So there was a conflict between the two psychologies. Once the rat tasted the sugar water again, it immediately became nauseous, and quickly lost the goal of ingesting it. Through the pain of nausea, the two psychologies had become "aligned."

The crucial point here, from my perspective, is that pleasure need not regulate human behavior by being part of the content of human desires. It need not be the thing we are explicitly aiming for. Rather, *the function of pleasure is to provide a kind of reinforcement mechanism that strengthens or weakens the values associated with certain goals*. Dickinson's work shows that pleasure need not be part of the content of a desire in order to serve as a reinforcement mechanism for our desires. A useful diagram in Dickinson (2008, 280) makes clear that he does not think pleasure is typically part of the representational content of ultimate desires. Rather, creatures typically have ultimate desires for things like food, water, and so on; when those desires are associated with pain, they are abolished; when they are associated with pleasure, they are reinforced. In this way, it seems to me that empirical work in the neuroscience of pleasure supports R-hedonism.

One could argue that, though the empirical work outlined here is *consistent* with R-hedonism, it is also consistent with I-hedonism. In other words, Dickinson's work does not support R-hedonism *over* I-hedonism.[6] It merely shows that there is some intimate

[6] I owe this observation to Chandra Sripada.

relationship between our desires and our hedonic states. I agree with that assessment. I have no doubt that the I-hedonist could give a natural and plausible interpretation of Dickinson's work. But that is not the point I am trying to establish here. The point of the last section is that we have good reason for rejecting I-hedonism. If Dickinson's HIT can be taken to support at least one of I-hedonism or R-hedonism, and if there are independent grounds to think that I-hedonism is false, then HIT can be taken to support R-hedonism.

Section 5. Four Objections.

I'll consider four objections to R-hedonism before closing. First, Section 3 discussed Sober and Wilson's evolutionary argument against I-hedonism. One might think that Sober and Wilson's argument against I-hedonism could be converted, fairly straightforwardly, into an argument against R-hedonism. Second, one might wonder if R-hedonism even deserves to be described as a type of "hedonism" at all, given that it is neutral about the contents of ultimate desires. Third, one might worry that R-hedonism is a semantically trivial doctrine. Fourth, why should one think that pleasure is the *only* mechanism by which ultimate desires are reinforced? In other words, one might grant that it constitutes *one* such mechanism. But R-hedonism aspires to a more exclusivist claim. How can this be defended? I'll take each of these in turn.

First, let's consider whether Sober and Wilson's (1998) argument against I-hedonism can be converted, fairly easily, into an argument against R-hedonism. Recall that the purpose of their argument against I-hedonism was to establish that altruism is more likely to have evolved by natural selection than I-hedonism. (Specifically, they are considering which of the two was more likely to have evolved as a mechanism for child-rearing, that is, as a mechanism for motivating parents to care for their children.) They argue that three criteria must be considered in order to decide which of the two "design solutions" was more likely. The first is *availability*: is there reason to think that the requisite variation (e.g., the right sorts of genetic mutations) would have existed in ancestral populations? In this respect, they argue, altruism and I-hedonism are comparable. The second is *reliability*: which of the two would make for a more reliable caretaker? Here, they argue that the altruistic solution is a more reliable mechanism for getting parents to take care of their children. The hedonist solution introduces a source of fragility that altruism avoids: namely, the hedonist will lose the desire to care for the children if that hedonist loses the belief that doing so is pleasurable.

The third criterion is *energetic efficiency*: simply put, is there any reason to think that one solution requires more calories to maintain than the other? Crucially, they argue that the two solutions are comparable here. After all, they both involve the same basic psychological *machinery*, namely, a belief-desire psychology that causes behavior. They only differ with respect to the *contents* of the desires they postulate. But it is hard to see why merely having a desire with one content (e.g., *you eat ice cream*) would be more metabolically costly than having a desire with some other content (e.g., *I eat ice cream*). But, as I will shortly explain, this is precisely what we *cannot* say about the relationship between R-hedonism and its alternatives. Hence, Sober and Wilson's powerful argument against I-hedonism cannot be converted, in any straightforward way, into an argument against R-hedonism. I will explain this point below.

Consider what would have to be established to convert this argument against I-hedonism into a parallel argument against R-hedonism. Keep in mind that the relevant comparison, in this case, is not between R-hedonism and altruism. Rather, it is between R-hedonism and *some other mechanism* the function of which is to maintain or reinforce desires in one's cognitive system. Presumably, this hypothetical alternate mechanism can be described in the following way: it causes the reinforcement of certain ultimate desires over others, but not, or not merely, by exploiting the connection between the individual's desires and the individual's hedonic states. I will refer to this alternative doctrine as NR-hedonism. In sum, here are the two theories we would like to compare:

> R-hedonism: desires are reinforced in one's cognitive system only by virtue of their association with hedonic states
>
> NR-hedonism: desires are reinforced in one's cognitive system but not (merely) by virtue of their association with hedonic states

We can now ask: *of* R-hedonism and NR-hedonism, which, if either, is more likely to have evolved?

Sober and Wilson's argument relies crucially on the observation that the only difference between altruism and I-hedonism is a difference with respect to the contents of the desires they postulate and *not* with respect to the underlying "machinery" of motivation. This allows them to argue that the two solutions are comparable with respect to energetic efficiency. But we cannot say this about the relation between R-hedonism and NR-hedonism, since the two theories are neutral about the contents of our ultimate desires. They differ precisely with respect to the basic psychological and neurological machinery by which desires are reinforced. As a consequence, we cannot claim that NR-hedonism (weakly) dominates R-hedonism in the absence of further empirical research. Sober and Wilson's argument cannot be used, in any obvious way, as a template for a corresponding argument against R-hedonism.

A second objection does not question the truth of R-hedonism, but the appropriateness of the label itself. If R-hedonism is not a theory about the contents of our ultimate desires, why call it a form of "hedonism" at all? Isn't this just abusing the label? It seems to me that it is not abusing the label. I'll first put the argument in a somewhat rough, informal way, and then in a more precise way. Roughly speaking, a core intuition that hedonists share is that pleasure is somehow "behind" (or, if one prefers, "above") all of our motives. In other words, many of the traditional debates are framed by a kind of spatial metaphor. Whether pleasure is "behind" our desires in the sense of being part of the content of our ultimate desires, or whether it is "behind" our desires in the sense of being a reinforcement mechanism for our desires, doesn't seem to me to be the crucial difference regarding the appropriateness of the label. As Bentham put it, using an even more colorful metaphor, human beings are under the "governance" of pain and pleasure.

I-hedonism and R-hedonism would just represent two alternative ways of "being under the governance" of pleasure.

Another way of putting the point, more precisely, is this: what is essential to hedonism is the claim that our ability to have desires regarding the welfare of others depends on their effect, or their presumed effect, on our hedonic states. If that link, or that presumed link, between the desire and the hedonic state, were broken, then the desire would be abolished. If one's theory of motivation has that implication, it seems to me that it would be appropriate to call it a "hedonistic" theory of motivation. But this is what R-hedonism entails.

A third sort of objection stems from a worry that R-hedonism is perhaps a semantically trivial doctrine.[7] The idea is something like this. R-hedonism states that ultimate desires are reinforced by virtue of their being associated with pleasure. But one might worry that R-hedonists will just use the term "pleasure" to denote *whatever mechanism it is that happens to reinforce ultimate desires*. This objection parallels an objection that has often been raised against I-hedonism. The objection is that the I-hedonist can use "pleasure" so loosely that it can refer to anything that people desire in an ultimate way, and hence I-hedonism is trivially true.

I agree that this objection has some merit. This is because, as I noted above, I do not have any very precise definition of "pleasure" to fall back on. I think the definition of pleasure is one that must be explicated by ongoing empirical research. For example, neuroscientists have only recently been able to distinguish sharply between "liking" and "wanting" on biological grounds (e.g., Berridge 2010) and I believe that this work is relevant to actually explicating the very meaning of "pleasure" and not just illuminating its biological foundations.

But I still do not think that R-hedonism is a tautology. It strikes me as entirely possible, as a *conceptual* matter, that an ultimate desire can be reinforced independently of its association with pleasure. Whether ultimate desires ever are, *in fact*, reinforced this way, is an empirical matter. For example, Berridge and O'Doherty (2013, 344; also see Robinson and Berridge 2013) describe how drug addiction can result from the impact of drugs on the mesolimbic dopamine tract, which is more closely implicated in wanting than pleasure. Berridge's hypothesis here is that repeated consumption of drugs alters the mesolimbic dopamine tract in such a way that it becomes hypersensitized to drugs. That is, drugs will always retain a special motivational salience ("wanting") for the addict, long after the pleasure of using drugs has diminished. This work suggests the *conceptual* possibility that a desire for a drug might be reinforced by a mechanism that has nothing to do with its (presumed) hedonic impact. Again, nobody has succeeded in bringing about such a total dissociation between "wanting" and anticipated "liking" in a human being. However, the fact that it is conceptually possible to do so would seem to show that R-hedonism is not a tautology.

[7] I thank Armin Schulz for raising this objection

The final objection is probably the strongest one. One might accept, on the basis of empirical considerations, that pleasure and pain perform *some* role in reinforcing our ultimate desires. But why should it be the *only* such mechanism? In other words, why would one want to exclude the possibility that an ultimate desire could be reinforced solely because of some other consequence that it has, quite independently of pleasure and pain? It is easy enough to accept that if the satisfaction of a certain ultimate desire, say, a desire to assist the homeless, were repeatedly associated with pain, then one might stop having that desire. It is also easy to accept that if the satisfaction of a desire to assist the homeless were repeatedly associated with pleasurable consequences, then one would maintain that desire, or that desire would grow even stronger. Pleasure and pain do seem to play a role in reinforcing our desires. But there could be some other mechanism that also plays a role in maintaining or reinforcing certain desires, and this mechanism could, in principle, operate independently of the first. Why accept a philosophical position that precludes this possibility in advance?

I certainly do not possess an argument that such a mechanism does not or cannot exist. This strikes me as an empirical question. In a sense, the dialectic here reflects a dialectic that has been going on for centuries. The traditional I-hedonist claims that people only have ultimate, self-directed desires; the traditional altruist grants that people do have such desires, but asserts that, in addition to those, people have ultimate, other-directed desires, too. A major strength of R-hedonism is that it grants to the hedonist his or her hedonistic intuitions, but avoids the major pitfalls of I-hedonism. In this sense, R-hedonism borrows its strength from the sorts of intuitions that have always sustained I-hedonism, and avoids all of its weaknesses. Hence, the traditional hedonist should see R-hedonism as a welcome advance.

Section 6. Conclusion

In the foregoing I have developed a distinction between two forms of psychological hedonism. The first, inferential hedonism, claims that people only have ultimate desires regarding their own hedonic states (pleasure and pain). The second, reinforcement hedonism, claims that, whatever the contents of one's ultimate desires, those desires are only reinforced by virtue of their association with pleasure. The first theory is about the content of one's ultimate desires; the second is about the mechanism by which those desires are reinforced. The reason that R-hedonism should be classified as a "form" of hedonism at all is because it represents one way that humans can be said to be "under the governance" of pain and pleasure. R-hedonism borrows the strengths of I-hedonism while avoiding its major pitfalls. This alone should make it a welcome advance for those who share hedonistic intuitions but who question the merits of I-hedonism on psychological or evolutionary grounds.

In addition to its interest for philosophers, the distinction between I-hedonism and R-hedonism represents a potentially important development for psychologists who wish to explore the relevance of recent neuroscience for traditional philosophical problems of the mind and morality (e.g., Schroeder 2004; Schroeder et al. 2010; Holton and Berridge forthcoming). By the same token, it might provide a conceptual tool for neuroscientists who are probing the nature of pleasure. For example, two neuroscientists, Siri Leknes and Irene Tracey, quote Bentham approvingly in a recent review of the neuroscience of pleasure (Leknes and Tracey 2008, 314). In doing so, they clearly recognize that traditional philosophical discussions about the nature of pleasure and its role in cognition are relevant to thinking about the neuroscience of pleasure. Some neuroscientists even refer to this branch of neuroscience as "hedonics," and the neural foundations of pleasure, "hedonic hotspots" (Smith et al. 2010; Dickinson and Balleine 2010). It seems to me that careful analysis of the different forms of psychological hedonism can provide a useful framework for this endeavor.

References

Bentham, J. 1789 (1780). *An Introduction to the Principles of Morals and Legislation*. Oxford: Clarendon Press.

Berridge, K. C. 2007. The debate over dopamine's role in reward: the case for incentive salience. *Psychopharmacology* 191: 391-431.

Berridge, K. C., and O'Doherty, J. P. 2013. From experienced utility to decision utility. In Glimcher, P. W., and Fehr, E. (Eds.) *Neuroeconomics: Decision Making and the Brain, 2nd ed.* Amsterdam: Elsevier, 335-351.

Butler, J. 1729. *Fifteen Sermons Preached at the Rolls Chapel*. London: Botham.

Dickinson, A. 2008. Why a rat is not a beast machine. In Weiskrantz, L., and Davies, M. (Eds.) *Frontiers of Consciousness: The Chichele Lectures*. Oxford: Oxford University Press, 275-288.

Dickinson, A., and Balleine, B. 2010. Hedonics: The cognitive-motivational interface. In Kringelbach, M. L., and Berridge, K. C. (*Eds.*), *Pleasures of the Brain*. Oxford: Oxford University Press, pp. 74-84.

Feinberg, J. 1987. "Psychological Egoism." In Sher, G., (Ed.) *Moral Psychology*. San Diego: Harcourt, Brace Jovanovich, 1-15.

Garson, J. 2011. Selected effects and causal role functions in the brain: The case for an etiological approach to neuroscience. *Biology and Philosophy* 26: 547-565.

Garson, J. 2012. Function, selection, and construction in the brain. *Synthese* 189: 451-481.

Garson, J. 2015. *The Biological Mind: A Philosophical Introduction*. London: Routledge.

Holton, R., and Berridge, K. Forthcoming. Addiction between compulsion and choice. In Levy, N. (Ed.), *Addiction and Self-Control*. Oxford: Oxford University Press.

Kringelbach, M. L., and Berridge, K. C. (*Eds.*), *Pleasures of the Brain*. Oxford: Oxford University Press

LaFollette, H. 1988. The truth in psychological egoism. In Feinberg, J. (*Ed.*), *Reason and Responsibility*. Belmont, CA: Wadsworth, pp. 500-507.

Millikan, R. G. 1984. *Language, Thought and Other Biological Categories*. Cambridge, MA: MIT Press.

Nozick, R. 1974. *Anarchy, State, and Utopia*. New York: Basic Books.

Robinson, M. J. F., and Berridge, K. C. 2013. Instant Transformation of Learned Repulsion into Motivational "Wanting." *Current Biology* 23: 282-289.

Sarkar, S., and Garson, J. 2004. Multiple criterion synchronization for conservation area network design: The use of non-dominated alternative sets. *Conservation and Society* 2:433-448.

Schroeder, T. 2004. *Three Faces of Desire*. Oxford: Oxford University Press.

Schroeder, T., Roskies, A. L., and Nichols, S. 2010. Moral motivation. In Doris, J. M. and the Moral Psychology Research Group (*Ed.*), *The Moral Psychology Handbook*. Oxford: Oxford University Press, pp. 72-110.

Schulz, AW. 2011. Sober and Wilson's Evolutionary Arguments for Psychological Altruism: A Reassessment. *Biology and Philosophy* 26: 251-260.

Smith, K. S., Malder, S. V., Peciña, S., and Berridge, K. C. Hedonic Hotspots: Generating Sensory Pleasure in the Brain. In Kringelbach, M. L., and Berridge, K. C. (*Eds.*), *Pleasures of the Brain*. Oxford: Oxford University Press, pp. 27-49.

Sober, E. 1992. Hedonism and Butler's Stone. *Ethics* 103: 97-103.

Sober, E. 1994. Did evolution make us psychological egoists? In Sober, E. (*Ed.*), *From a Biological Point of View: Essays in Evolutionary Philosophy*. New York: Cambridge University Press, pp. 8-27.

Sober, E., & Wilson, D. S. 1998. *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Cambridge, MA: Harvard University Press.

Sterelny, K. 2003. *Thought in a Hostile World: The Evolution of Human Cognition*. Malden, MA: Blackwell.

Stich, S. 1978. Beliefs and Subdoxastic States. *Philosophy of Science* 45: 499-518.

Stich, S. 2007. *Evolution, Altruism, and Cognitive Architecture: A Critique of Sober and Wilson's Argument*. Biology and Philosophy 22: 267-281.

Stich, S., Doris, J. M., and Roedder, E. 2010. Altruism. In Doris, J. M. and the Moral Psychology Research Group (*Ed.*), *The Moral Psychology Handbook*. Oxford: Oxford University Press, pp. 147-205.