The Journal of Mind and Behavior
Spring 2017, Volume 38, Number 2
Pages 91–110
ISSN 0271–0137

Strawson's Case for Mental Passivity

George Seli

St. John's University

Galen Strawson (2003) argues that relatively few of our mental events are actions, what I refer to as the non-agentive thought thesis (NATT). NATT restricts the actional kinds of mental event to volition and catalysis, the latter being the mental preparation to receive thoughts. Strawson supports NATT on both metaphysical and phenomenological grounds. Metaphysically, his primary argument is that the results of thinking — whether decisions, judgments, creative ideas, etc. — are not "intentionally controlled," which disqualifies them as actions. Phenomenologically, Strawson claims that, at least for him, the sense of mental agency is limited to volition and catalysis, and he takes this passivist phenomenology to further support NATT. I raise problems for both arguments. On the metaphysical side, I argue that Strawson's understanding of intentional control as direct control yields an unreasonable constraint on actional mental events. The phenomenological argument is undercut by the empirical possibility that Strawson's passivist phenomenology is cognitively penetrated by his own anti-agency beliefs.

Keywords: mental agency, mental action, cogntive penetration, phenomenology

In contrast to Peacocke (2007), who claims that "Much conscious thought consists of mental actions" (p. 358), Strawson (2003) argues that mental agency is very limited in scope, which is to say that relatively few mental events are actions. I will refer to Strawson's view as the *non-agentive thought thesis* (NATT). On this view, mental action is restricted to volition and the preparation to receive thought. The latter he refers to as "catalytic" or "priming" activity, and it includes phenomena Strawson describes as "setting one's mind at the problem" one wants to solve, "focused concentration of will," a "receptive blanking of the mind," "maintaining attention," and so forth (pp. 236–237). Then, if all goes well cognitively, the "content outcomes are delivered into consciousness" by the "natural causality of reason," or by imaginative/associative processes, as the case may be. Per NATT, neither these processes nor their outcomes are actions. Thus, while Peacocke (2007, sect. II) maintains that deciding, judging, accepting,

Correspondence concerning this article should be addressed to George Seli, Philosophy Department, St. John's University, 8000 Utopia Parkway, Queens, New York 11439. Email: selig@stjohns.edu

calculating, reasoning, and trying are all actions, Strawson would say that only the last one in this list, being catalytic in nature, is an action.[1]

In support of NATT, Strawson offers both a metaphysical and a phenomenological line of argument. Metaphysically, he observes that cognitive processes such as deliberation, judging, and associative thinking are "ballistic," that is, they operate largely independently of our will (involuntarily) once initiated. He then argues that the result of such a process — whether a decision, judgment, creative idea, etc. — is not "intentionally controlled," which disqualifies the outcome as an action. Phenomenologically, Strawson claims that, at least for him, the *sense* of mental agency is quite limited in scope: only volition and catalysis feel agentive. This passivist phenomenology is intended as further support to the claim that only those mental events are actions.

This paper will challenge both arguments. On the metaphysical side, I argue that Strawson's understanding of intentional control as *direct* control yields an unreasonable constraint on actional mental events. The phenomenological argument is undercut by the empirical possibility that Strawson's phenomenology is *cognitively penetrated* by his own anti-agency beliefs. As such, the passivist phenomenology would not offer independent support for NATT.

I begin with a discussion of Strawson's phenomenological grounds for the thesis. First, I offer empirical support for the idea that felt agency can be cognitively penetrated, and argue that in the case of a passivist phenomenology of mental action, the sense of agency *is* being cognitively penetrated. Second, I outline what I take to be the natural (i.e., non-penetrated) alternative to the passivist phenomenology, namely a phenomenology in which many kinds of mental events do feel agentive to some degree. Third, I present Strawson's phenomenology in more detail, including the types of mental events that passivism extends to and the subjective awareness of the ballistic process leading to those events. Fourth, I contend that his passivist experience is produced by beliefs that (i) divorce the ballistic process from the self, (ii) discredit the role of volition and catalysis in initiating the process, and (iii) deny that we have intentional control over the process's outcome.

Among these beliefs, Strawson is most clearly committed to (iii), insofar as his metaphysical argument is based on it. Toward evaluating this argument, I examine several understandings of "intentional control" based on passages in Strawson (2003). The interpretation that would most strongly support NATT is the claim that we lack direct control over content outcomes; they are achieved "by doing something else." I counterargue that requiring actional mental events to be directly controlled is an unreasonable constraint. Moreover, I argue that we lack

[1] I understand trying as a distinct type of mental event, as Peacocke does, and take it to be a case of Strawsonian catalysis: exerting mental effort to perform an act. Alternatively, trying can be understood as a set of events (a "composite" event) including reasons for acting and a resulting intention to do so (see Adams, 1997, 289–291).

such control over the kinds of mental event Strawson *does* take to be actions. I conclude with a brief discussion of the methodology for investigating the scope of mental agency.

Strawson's Phenomenological Argument

Strawson's phenomenological argument is presented less explicitly than his metaphysical argument, so it is worth highlighting a few passages where phenomenological support for NATT is given or at least implied. I discuss them in turn.

(1) *"When I consider my mental life I find that things constantly impinge on me. I remember that I have to do X — it strikes me that Y is true. I want some coffee — I wonder where the filter papers are. I know I have to go to Charing Cross — I find myself thinking about the best way to travel. Thought, it seems, is often a matter of things just happening"* (2003 p. 229). This passage concludes that such thoughts "just happen" — i.e., are not intentionally controlled — based on various premises describing how they *feel* like they just happen. Note that Strawson is citing cases of thoughts that just "pop into mind" unwilled, where most people would likely *not* feel very agentive. So the cases are not the best phenomenological evidence for NATT, which denies that even *deliberate* cognition is agentive. The following quoted passages concern cases of phenomenal passivity in intentional reasoning, which is better support for the view.

(2) *"No doubt there are … preparatory, ground-setting, tuning, retuning, shepherding, active moves or intentional initiations. But action, in thinking, really goes no further than this. The rest is waiting, seeing if anything happens, waiting for content to come to mind, for the 'natural causality of reason' to operate in one"* (p. 232). *"At other times there is a deliberate setting of the mind at the problem of what to do, a process of focusing on the problem, a concertion of thought, and this can be a matter of action. But what follows is, again, just a waiting for content to occur"* (p. 243). In these passages, NATT is supported by citing an *introspective experience of non-action*: waiting for content. To Strawson it does not feel as if he is bringing about the content; rather, it feels like he is waiting for it. So, the content's arrival must be nonagentive.

(3) *" 'Every kind of reasoning is nothing, in its simplest form, but attention,' as Shadworth Hodgson remarked. It is a laying open of oneself to the 'natural causality of reason,' an induction of oneself into a receptive, actively passive state, tuned this way or that"* (p. 238). *"When one sets oneself to imagine anything there comes a moment when what one does is precisely to relinquish control"* (pp. 242–243). These passages cite an *experience of relinquishing control*, and the implied conclusion is that one is not in control in seeking a content outcome. Thus, we have another

case of phenomenological support for NATT. If Strawson is not referring to the *experience of* relinquishing control here, but rather citing the putative fact that we *do* relinquish control, then he has question-begging support for NATT.

The subjective reports that ground Strawson's phenomenological argument are of course difficult to discredit, but my objective is not to cast doubt on his passivist phenomenology. Rather, I hold that it is unreliable support for NATT, given the possibility that the phenomenology is cognitively penetrated by Strawson's own beliefs about mental agency.

If the feeling of agency were penetrated by views on the nature of mental action, the resulting phenomenology would not offer *independent support* for a metaphysical thesis such as NATT. The penetrating beliefs support mental passivity, so of course they would engender a passivist phenomenology — one that therefore does no more to support NATT than the beliefs themselves. Similarly, if one's perceptual experience were penetrated by one's belief that *P* such that the experience makes it seem that *P* is the case, that experience would arguably fail to independently support *P*. Put differently, the experience would fail to "epistemically elevate" one's belief that *P* (see, e.g., Siegel, 2012, on this point).

Evidence for the Penetrability of the Sense of Agency

Clearly, the forgoing argument depends on the premise that the sense of mental agency — like perceptual experience — *can* be cognitively penetrated by beliefs; that is to say, its phenomenal character can be modulated by those states. First, I would note that the sense of agency in general is typically "thin" and "evasive" as compared to perceptual qualia (Metzinger, 2004), and "minimal" when one is immersed in action (Marcel, 2003), whether mental or bodily. As such, one can expect it to be especially susceptible to being disrupted by belief possession. Second, the penetrability hypothesis has some empirical support from experiments on the influence of beliefs on agency feelings.

One well-known example in the literature is the *I Spy* study (Wegner and Wheatley, 1999; reviewed in Wegner, 2002, p. 74), which provides evidence that agency feelings can be increased by beliefs about the close temporal priority of a thought about a behavior to the behavior itself. Two participants, a subject and a confederate, jointly moved a cursor around a computer screen displaying about 50 items. Per the instructions, they were to move the cursor around at will for about 30 seconds while hearing music and occasional distracting words via headphones. Each time the cursor was placed on an item, the participants independently had to rate the degree to which the stop was intended on a 0-100 scale. Some stops were forced by the confederate, via instructions received on her headphones alone. On these, the subject heard the name of the target object via headphones either thirty, five, or one second before the stop, or one second after it. It was found that the more proximately the name was heard *prior* to the

stop, the more intended the subject rated the stop, even though it was forced. First, it is plausible that upon these cursor stops, the subject realized how closely her thought about the item preceded the stop on the item — call this kind of realization her *temporal precedence belief*. Second, the subject arguably bases her intentionality rating on her *feeling of agency*, or something phenomenologically akin to that feeling. Wegner seems to assume this: "there was an increased *experience of intention* when the thought was primed 1 to 5 seconds before the forced action" (p. 77; emphasis added). The subject is motivated to monitor her agency feelings, as the experimenter at the outset informed the participants that "the study was to investigate people's feelings of intentions for acts" (p. 74). Moreover, a judgment of intentionality that is made in a *graded* way (how *much* a given stop is intended) is likely based on a feeling, as feelings admit of degrees.[2] Thus, the study implies a correlation between temporal precedence beliefs and feelings of agency over cursor stops, namely, temporal precedence beliefs for more proximate stops followed by stronger feelings of agency for those stops. This is evidence that the temporal precedence beliefs are causing those feelings. And if that is possible, surely beliefs can *modulate* existing feelings of agency.

More recently, Desantis, Roussel, and Waszak (2011) suggested causal beliefs can promote intentional binding, i.e., the perceived shorter time interval between a voluntary action and its effect (see Haggard, Clark, and Kalogeras, 2002). Desantis et al. found that subjects intentionally bound their action to a tone when they were induced to believe that they would produce that tone, as opposed to another participant or an unclear source. Haggard et al. (2002) proposed that the intentional binding phenomenon they discovered engenders the sense of agency: "These results suggest that the brain contains a specific cognitive module that binds intentional actions to their effects to construct a coherent conscious experience of our own agency" (p. 385). But alternatively, the converse may be true: the sense of agency may be the proximate cause of intentional binding. Either way, the result of Desantis et al. (2011) is evidence that a causal belief can influence the sense of agency: the belief that one is the cause of the tone can promote that feeling *via* promoting intentional binding, or alternatively, it can promote that feeling *directly*, in turn creating intentional binding.

It has been proposed (Synofzik, Vosgerau, and Newen, 2008) that the sense of agency has both implicit (i.e., pre-reflective, nonconceptual) and explicit (reflective, conceptual) components. Intentional binding is thought to be associated with the implicit component; that is, the binding of actions to their sensory effects correlates with the agency the subject *feels*, as opposed to the agency she *judges* herself to have. Exploiting this connection, Lynn, Muhle–Karbe, Aarts, and Brass

[2] Experimental work based on felt agency supports the notion that the feeling admits of degrees. See, e.g., Farrer, Franck, Georgieff, Frith, Decety, and Jeannerod (2003) and Synofzik, Vosgerau, and Newen (2008).

(2014) provide evidence that anti-free will beliefs can *diminish* the implicit sense of agency by showing that such belief possession reduces intentional binding. Prior to engaging in an intentional binding task, participants read essays that promoted disbelief in free will along with general readings on consciousness, with the latter serving as a control. Participants were told they would be tested on the material with the primary goal of determining retention, in order to conceal the purpose of the experiment. Ultimately, Lynn et al. found that "intentional binding was significantly reduced in the anti-free will condition, indicating that determinist beliefs hamper the implicit sensation of being in control of one's actions" (p. 6). This result, though based on a test that involves *physical* actions, lends more credibility to my premise that certain beliefs about mental agency can reduce the feeling of control over one's cognition.

Now, there is significant support (e.g., Frith, Blakemore, and Wolpert, 2000) for the position that the sense of agency depends on a subpersonal "comparator" mechanism, as opposed to beliefs and inferences. This system generates a representation of the predicted sensory consequences of a movement based on a copy of the motor commands, and compares that representation with a representation of the actual sensory consequences. A "match" results in a feeling of agency, ex hypothesi. And as Feinberg (1978) originally proposed, such a subpersonal system may also ground the sense of mental agency, if thoughts are motor processes.[3] However, a subpersonal-level explanation of the sense of mental agency is compatible with my premise, which is that beliefs can modulate the feeling. That leaves open the possibility that the feeling is grounded in a motor process, such as a comparator for mental acts.

Assuming that such modulation can occur, a further question may be asked: Why think that cognitive penetration *is* occurring in the case of a passivist phenomenology of mental action, such as Strawson's? As I discuss in the next section, a perceived match between the content of a volition and a subsequent mental event supports the feeling of agency. If that is so, there must be some other psychological factor *counteracting* the effect of perceiving a match if a non-agentive feeling is the norm. I will argue that such a factor is a penetrating anti-agency belief.

A "Balanced" Phenomenology of Mental Agency

Prior to discussing the sources of a passivist phenomenology of mental action, it will be useful to contrast that phenomenology to a more "balanced" one, where some mental events do not feel like things we do (they seem to "just happen"), others feel somewhat under our control, and others feel fully like things we do.

[3] One questionable implication of this view, discussed by Campbell (1999, p. 618), is that the subpersonal comparator for thought would need to be sensitive to the semantic properties of propositional attitudes in order to be able to confirm that a given thought is the one the agent intended to think.

Following is a suggested taxonomy of mental event types that is based on whether they typically feel non-agentive (i.e., passive), semi-agentive, or fully agentive. Each list is clearly not exhaustive and intended only to provide examples.

(a) *Phenomenally non-agentive*: feelings, sensations, doubts, worries, urges, non-directed versions[4] of the mental events in categories (b) and (c)

(b) *Phenomenally semi-agentive*: the results of reasoning (e.g., judgments) and associative thinking (e.g., creative ideas), remembering

(c) *Phenomenally fully agentive*: mental-image making, supposing, subvocal speech

Admittedly, some kinds of mental events cannot easily be categorized on this scheme. For example, a change in one's mood may feel nonagentive, or it may feel agentive to some degree, if it was brought about deliberately; it does not seem to "typically" feel either way.

For (b) and (c) events, I assume that what underlies — or at least contributes to — felt agency is that they are brought about (in part) by volitions; that is, they are cases of *directed cognition*. I equate "volition" with Pacherie's (2007) "present-directed intention": a P-intention "anchor[s] the action plan both in time and in the situation of action" (p. 3). As such, a volition is an immediate mental precursor to a specific action, along with catalysis. This etiology enables the detection of a *match* between the content of a volition and a subsequent mental event. Specifically, the representational content of a volition to perform a certain mental action would provide the satisfaction conditions. Catalytic acts — e.g., "setting one's mind at a problem," "focused concentration of will," a "receptive blanking of the mind," "maintaining attention" — are kinds of mental effort that follow and help to carry out a volition to perform a certain mental act; they generally do not have content with which an outcome can match. For instance, I have a volition to decide what to have for dinner, and that volition is followed by a receptivity to ideas. I subsequently think *I'll have pasta*, a thought that matches the content of my will, but not the content of the state of receptivity (which is likely a nonrepresentational state).

The view that a match with volitional content underlies felt agency is both intuitively correct and well supported in the literature on the sense of agency. For example, Wegner (2002) argues that a judgment of the "consistency" between act and volition supports that sense. Pacherie (2007) considers the perceived match as the basis of the sense of control, which contributes to the sense of agency: "At the level of P-intentions [present-directed intentions], the sense that one is in control would rely on the perceived match or mismatch between the predicted perceptual effects, corresponding to the situated goal, and the actual perceptual

[4] These would be cases when decisions, judgments, recollections, mental images, etc. just "come to us." Or, in the case of judgments and decisions, they may come via reasoning, but the reasoning isn't catalyzed and engages automatically (see Buckareff, 2005, p. 85).

effects" (p. 19). There are, of course, no predicted *perceptual* effects of a mental act, but rather a predicted act whose occurrence can be introspectively confirmed. Now, it may be held that the "matching" instead occurs at the subpersonal level via a comparator mechanism (as discussed in the previous section). But this subpersonal basis for the sense of agency is increasingly regarded as *one among several factors* that can explain the phenomenon (Schlosser, 2012, pp. 143–144); personal-level matching would thus not be preempted as a source of the feeling. And second, it is questionable whether there is a comparator for mental acts. So, the personal-level matching hypothesis is strengthened for these acts.[5]

As I argue below, the reason for the lack of a *full* sense of agency for (b) events is one's awareness of a cognitive process *apart* from volition and catalysis in helping to generate the events. Type (a) events, on the other hand, are typically not willed (although feelings and sensations are sometimes brought about deliberately). Ex hypothesi, they typically feel non-agentive, as there is no volition whose content they can match.

I should point out that matching the content of a volition cannot be *necessary* for a mental event to feel agentive. For a volition itself feels agentive (when it is conscious), and it cannot feel that way due to perceived consistency with a prior volition on pain of infinite regress. I suggest that the reason volitions belong in the (c) category is because they are perceived — typically nonconsciously — to be *in accord with beliefs and desires*. My volition and subsequent effort to solve a problem, for instance, are in accord with my belief that solving the problem is important, my desire to get the answer, my belief that now is a good time to try to solve it, etc. This perceived consistency between the volition and one's belief/desire complex plausibly promotes an agentive feeling about the event.

Awareness of a Content Delivery Mechanism

For Strawson (2003), mental event types (b) and (c) generally do *not* feel like actions (to any degree). As noted previously, one aspect of his phenomenology is the feeling of "waiting for content to occur" (p. 243). There can indeed be an experienced time lag between catalysis and content outcome, particularly where the outcome results from reasoning, associative thinking, or remembering — the

[5] I do not think that the detection of a match need be a conscious judgment of consistency between volition and act, or require that either event be conscious. Wegner appears to assume this: "It is only when a thought is conscious prior to action that it can enter into the person's interpretation of personal agency and so influence the person's experience of will" (2002, p. 164). But surely nonconscious mental states, such as nonconscious volitions, can influence conscious ones, such as a conscious sense of agency. Moreover, one's introspectively confirming that a certain mental event matches the content of one's volition need not be done by representing either mental event, which entails that the represented event is conscious on the higher-order theory of consciousness. I discuss this point in Seli (2012, pp. 309–310).

processes that lead to type (b) events. And while the experience of this time lag need not be an experience of passivity ("waiting"), there is certainly an awareness of the operation of the cognitive faculties that deliver the content.

So, for example, if I want to come up with the answer to a division problem such as 1216/8, I catalyze my arithmetic faculties which, in turn, begin to deliver content: I might first get the imperative thought *Divide 12 by 8*; I catalyze a bit more and receive *1 with a remainder of 4* … the process continues until a thought with the content 152 (the answer) is delivered. Another example: in 1960, a Boston resident wants to submit a name for the city's new football team; he catalyzes his faculty of associative thinking and soon thinks of Boston's role in American history, then receives the thought *American Revolution*, and finally *"Patriots"* is delivered to his consciousness.

Clearly, one is not aware of many or all of the process's stages; many of the inferential steps in reasoning, for example, are often nonconscious. But the time lag effectively makes one aware *that* the process is operating, if not *how* it operates. For example, if the thinking of a name occurs with some delay after the effort to recall it, during that delay one becomes aware that one's mnemonic faculty is operating, though not the nature of the process. In other cases, one may have in mind a specific mediating process that one wishes to initiate in order to achieve a cognitive goal. Mele (2009) distinguishes between *trying to x* and *trying to bring it about that one x-s*, where the latter involves deploying some process that (hopefully) will cause one to *x*. For example, in trying to recall what he had for dinner on a given day, Mele notes that he will try to bring about this recollection by asking himself (silently) what other things he did on that day (p. 19). Dorsch (2009) makes a similar point in discussing "the instrumental reliance on certain epistemic or merely causal processes and their passive effects" (p. 41); for example, the deliberate consideration of certain evidence that one hopes will lead one to judge correctly on a certain issue. Just as in the case of merely waiting for the arrival of the desired thought after priming, these deliberate uses of mediating cognition entail awareness that mediation is happening.

Note that even in cases where the cognitive result follows smoothly from catalysis and there is no awareness or deliberate use of a mediating process, Strawson (2003) draws our attention to the fact that some kind of ballistic process is at work. Consider mental-image making: here one tries to form an image and then "waits for the *mechanism of imagination* — the (involuntary) spontaneity of imagination — to deliver the image," he writes (p. 243; emphasis added). This is questionable; it may be that there is only waiting in the case of images that are intricate or hard to picture. And is there any waiting between catalysis and supposing, or engaging in subvocal speech? These mental events seem to occur instantaneously after one has the slightest volition or "inclining of the mind" to bring them about.

Nonetheless, I agree with Strawson that one is often aware of some mediating process that plays a causal role in reaching one's cognitive goal. I also think that

awareness of this content delivery mechanism might, for some people, *reduce* the feeling of agency over the outcome; I explain why in my discussion of the feeling of semi-agency below. But I doubt that said awareness can *eliminate* the felt agency that is promoted by a match with volitional content. Such awareness is common (occurring whenever one attempts a challenging cognitive task), but a passivist phenomenology of cognition is likely not. There must be an additional psychological element that gives rise to such a phenomenology, and I argue it is the possession of anti-agency beliefs.

Penetration by Anti-Agency Beliefs

Importantly, Strawson does not characterize his directed cognition merely as "passive," but also with locutions that imply certain beliefs about the nature of that cognition. In particular, "waiting for content to be delivered" by the ballistic process suggests he holds two implicit beliefs regarding volition *v*, act of catalysis *c*, mediating process *m* and cognitive result *r*:

(i) *m*, unlike *v* and *c*, is not part of the self
(ii) *m* causes *r*, but neither *v* nor *c* causes *r*

Regarding (i), suppose a person *self-identified* with *m*; that is, she believed the ballistic process to be partly constitutive of herself. She might then use a locution such as "waiting for my rational component to deliver *r*." And since the "delivery" of something to oneself is a rather peculiar usage, perhaps instead: "waiting for my rational component to *produce r*."

Why might one not self-identify with *m*? Perhaps in view of the following argument: only *conscious* mental events and processes constitute the self, and since much (or all) of *m* is nonconscious, the process is not part of the self. It would then be logical to describe one's phenomenology as *r* being *delivered* to consciousness (i.e., oneself), by a process that is *apart from oneself*. As I will discuss shortly, there is reason to attribute to Strawson the belief that nonconscious mental phenomena are not part of the self. But first it is important to observe two claims that he does *not* make regarding the nonconsciousness of the ballistic process: Strawson does not take that fact to establish NATT, nor to establish that nonconscious mental phenomena do not *belong to* the self. As to the first point, Strawson explains: "This non-consciousness is itself an important fact, I think, and invites reflection. Some may think that it amounts already to the point that the essence of thinking (as opposed to the supporting work of catalysis and priming) is not a matter of action. This may not be the right reaction, all things considered…. The coming to mind itself — the actual occurrence of thoughts, *conscious or non-conscious* — is not a matter of action" (p. 234; emphasis added). As to the second point, Strawson explains that "our thoughts and judgments are not in any sense not our own, or less our own, for not being direct products of

consciousness" (p. 247). This statement suggests he may also consider the nonconscious processes causing those products to belong to the self.

However, the belief that would support an agency feeling is not merely that one's "ballistic machinery" is *one's own*, but that it is (partly) *constitutive of oneself*. After all, one owns many things that do not plausibly constitute one's self, a library card, a car, etc. The belief, for example, that the mathematical cogitations that mediate between my willingness to solve problem X and my thinking of the solution to X are not merely "mine," but *me*, supports the feeling that *I* have solved X. Indeed, the sense of agency is often defined as the feeling of "authoring" an action. So, Strawson may well hold (i); that is, he divorces *m* from the self.

Further support for this belief attribution comes from the concept of the self that Strawson has advanced in other writings. He has argued that there exists a kind of minimal self that is "a single mental thing that is a conscious subject of experience" (1997, p. 407). A nonconscious mental process, such as nonconscious calculation, is thus excluded from this self for several reasons: it is not conscious, nor a subject of experience, nor a thing (it's a *process*). Strawson (2009) calls this minimal self a SESMET: a **S**ubject of **E**xperience that is a **S**ingle **ME**ntal **T**hing. While he allows that the concept of the self typically encompasses other features, most notably one's agency as well as personality and persistence over time, it *need not*: one can coherently conceive of oneself simply as a single mental thing that has experiences. Furthermore, he argues that this concept is metaphysically accurate, as SESMETS do exist. Thus, not only does Strawson hold that agentive properties are not essential to the self; he also apparently holds that the processes that carry out our mental agency (the ballistic ones) are not even accidental constituents of the self — insofar as he does not cite them as potential components of the self-concept. Volitional and catalytic abilities, on the other hand, could at least be accidental properties of the self for Strawson, as he considers them under the scope of our agency.

Regarding (ii), if a person also believed that *v* and *c* are causes of *m*, and thereby *r*, the characterization of her experience might be further amended: "causing [instead of waiting for] my rational component to produce *r*." But while the phrasing of Strawson's introspective report suggests he holds (ii), he seems to allow that *c can* be a cause of *r*, per the following passage: "I also agree that the occurrence of our thoughts and choices can be *partly caused* by genuinely intentional mental actions on our part — the catalytic business … the girding of the mind to engage the problem at hand" (2003, p. 248; emphasis added).

Since Strawson appears to allow that catalysis is efficacious, it is doubtful that he would deny volition a causal role. We naturally think that our volitions to bring about our cognitive goals are efficacious,[6] and the view is quite defensible. Based

[6] Note that this view does not require that mental events be experienced as caused by intentions. Perhaps there is an "experience of intentional causation," as Pacherie (2007) contends. Then again, such events may be experienced simply as things we do (Schlosser, 2012).

on a reductionist theory, mental events such as volitions are efficacious insofar as they are numerically identical to efficacious neural events. Alternatively, based on a supervenience view, mental events may be efficacious as causally relevant determinables supervening on the neural events that realize them, as argued by Yablo (1992) toward solving the problem of the causal exclusion of the mental posed by Kim (1989). Now, it may be argued that conscious volitions *specifically* are inefficacious, as they are simply a byproduct of their neural antecedents (e.g., the readiness potential), which do the real work in driving action. This is Wegner's (2002) view, based on the results of Libet's (1983) seminal experiments, and presumably it would apply as well to conscious volitions to perform mental actions. However, it is quite plausible that a preceding *nonconscious* will is efficacious in that case (see Rosenthal, 2002).

Instead of (ii), Strawson's passivist introspective report ("waiting for content") may reflect a different sort of belief that could *also* induce a feeling of passivity over a content outcome; namely, that we lack "intentional control" over that outcome (2003, p. 234). That type of belief can be expressed as follows:

(iii) *v* does not control *r*

I will classify (ii) and (iii) as beliefs about *conative limitation*, as they both assert that something about the process of conation is ineffectual: either (ii) volition and catalysis are causally inert, or (iii) volitions do not control the outcomes they represent. If, as seems to be the case, Strawson holds (iii) but does not commit to (ii), his position would be that although volition and catalysis can be efficacious, the volition that precedes catalysis does not "control" the outcome or the ballistic process that yields it. (I analyze the idea of intentional control in my discussion of Strawson's metaphysical argument below.)

Consider his claim that ballistic processes are "spontaneous," which he defines as "involuntary, not due to conscious volition" (p. 232). Elsewhere he writes, "Call what goes on mental spontaneity if you like, allow the arising of contents to be a matter of spontaneity; but admit, then, that spontaneity has nothing particularly to do with action *or will*" (p. 233; emphasis added). If Strawson is not asserting that volition is simply inefficacious in these passages, then he is referring to some type of control volition lacks over content outcomes. And if one believes, implicitly or explicitly, that one's will fails to control the outcomes of reasoning, imagination, etc., one may well *feel* that one lacks control over those outcomes, per the hypothesis that agency feelings are cognitively penetrable.

I do not question the passive nature of Strawson's experience during directed cognition, and if he does not describe his experience in more agentive ways, it is surely because such descriptions would be inaccurate. My argument concerns the possibility that belief (i), together with a belief about conative limitation, penetrate his experience and render it passive. As the product of those beliefs, the phenomenology would not offer any more support for NATT than the beliefs themselves.

Explaining the Feeling of "Semi-Agency"

Before proceeding to examine Strawson's metaphysical argument for NATT, I wish to offer an explanation of how a feeling of semi-agency arises for type (b) mental events (discussed in the foregoing section, "A 'Balanced' Phenomenology of Mental Agency"). Let us suppose that an agent does not hold (i)–(iii) with regard to her own cognition. That is, she self-identifies with the ballistic processes that carry out her volition, considers her volition and catalytic acts efficacious, and considers the outcome to be controlled by her volition. Why might a sense of *semi*-agency result qua the outcome, as opposed to a full sense of agency? The reason, I suggest, lies in the awareness of the causal involvement of a content delivery mechanism for type (b) mental events (as discussed above). Due to that awareness, the agent will believe there is a cause of the cognitive goal state *apart from volition and catalysis*.

This belief might weaken the sense of agency, based on the exclusivity criterion. Part of Wegner's (2002) theory of apparent mental causation, the criterion entails that we would get the full sense of agency only if we judge that the act (in this case a mental act) has *no other causes but our volition*. In the case of mediated mental agency (where the mediation is apparent to the thinker), this criterion appears not to be met: my volition to solve a given math problem, for example, is independent of my mathematical faculties, which *also* cause my thinking the solution. Ex hypothesi, I should feel (at best) semi-agentive in generating that solution.

In discussing the exclusivity criterion, Wegner cites internal (mental) causes that may lead an agent to doubt whether her "thought" (i.e., volition) is the real cause of some action of hers. We may refer to these as *competing causes* — those that may in fact be the *sole* cause of the action. Competing internal causes, Wegner suggests, can include emotions, dispositions, habits, and impulses. For example, a person may feel agentive in the act of gambling, believing that the behavior is consistent with and caused by his will. But that sense will likely be weakened if he becomes aware that he has a gambling habit that may be the actual cause of his behavior, Wegner argues. "Whenever we become aware of some cause of our action that lies inside ourselves but of which we were previously unconscious, we may lose some sense of will," Wegner writes (2002, pp. 90–91). That is, a sense of semi-agency may result.

Using the forgoing notation, the question is, should m be considered an internal cause of r that *competes* with volition v, in the sense that m, and not v, may be the actual cause of r? While v is not the exclusive cause of r, neither does m compete with v, I argue. Since it is plausible that v causes m, we would have the causal chain $v \rightarrow m \rightarrow r$. In contrast, a person's will to gamble would not cause his gambling habit; indeed, the habit is suspected to cause his will. Thus, $v \rightarrow m \rightarrow r$ does not entail causal competition between v and m. But even if the agent does not perceive m as a causal competitor to her volition, she may realize that she

engages her faculties of reasoning, creativity/associative thinking, etc., but does not will every stage of their operation; their operation is thus *to an extent* autonomous. Similarly, she may lose *some* feeling of power over a television's coming on if she considers the workings of the remote control that she does not specifically will into operation.

Now, one might ask, why would that sort of awareness of *m* engender a feeling of semi-agency instead of full passivity? I argue that what preserves some sense of agency in such cases is the perceived match between one's volition to perform *r* and the occurrence of *r* (as discussed previously). The belief that one's volition controls the result (the denial of (iii), in the foregoing section, "Penetration by Anti-Agency Beliefs") — supported by the perceived match — would also promote that feeling. And in the case of a mediated *mental* act, there is an additional kind of belief that may support felt agency; namely, the belief that the mediating mental process — rational, epistemic, causal — is part of one's "self" (the denial of (i), above). One would not normally have this sort of belief with regard to the operations of an external device such as a remote control.

Strawson's Metaphysical Argument for NATT

Strawson bases his metaphysical argument on (iii), the idea that we lack intentional control over the outcomes of cognitive processes such as reasoning, associative thinking, and imagination. What exactly is the nature of this conative limitation? Based on certain key passages in his paper, it might be understood in one of the following four ways. I argue that based on these interpretations, either we do not in fact lack intentional control over content outcomes (1 and 2, below), or we do lack such control, but the conception of control imposes an unreasonable constraint (3, 4). The most defensible sense in which we lack intentional control is (4): we do not *directly* bring about a content outcome. Thus, I will devote a separate section to that interpretation.

(1) *v's control over r is merely its representing r*. In willing to think something, one represents the mental act one wants to perform. Using imagination as an example, Strawson suggests that such representation is *all there is* to intentionally controlling that event: "When one has set oneself to imagine something one must obviously start from some conceptual or linguistic specification of the content (*spangled pink elephant*), and given that one's imagining duly fits the specification one may say that it is intentionally produced," he writes. "*But there isn't intentional control in any further sense*: the rest is a matter of ballistics, mental ballistics" (2003, p. 243; second emphasis added). In fact, apart from specifying the nature of the intended cognitive event, there *is* intentional control in a further sense: the representation is *causally* relevant to the occurrence of the event. Indeed, without acknowledging that causal relevancy, Strawson would be committed to (ii). Moreover, he allows

that "one *must* obviously start from some conceptual or linguistic specification of the content" (p. 241; emphasis added) and that the cognitive outcome is intentionally "produced," which imply causal relevancy. And since causation by the intentional representation entails (some) control, it is false that one lacks intentional control in the sense of *merely representing* the outcome.

(2) *v is causally insufficient for r*. On this construal, *v* lacks intentional control in the sense that it is sufficient to cause "some content or other," but insufficient to cause *r* with its particular content. Strawson writes: "And now I'm going to think something — I don't yet know what — and my thinking it is going to be a premeditated action: *swifts live their lives on the wing*.... In [this] case ... there is again a certain sort of action: an action of setting oneself to produce some content or other. But what happens then is — a content just comes. *Which particular content it is is not intentionally controlled*" (p. 239; emphasis added). However, on the assumption that there is the causal chain $v \rightarrow m \rightarrow r$, where *v* is causally sufficient for *m* to initiate, and *m* is causally sufficient for *r*, *v* is indeed causally sufficient for *r* (due to the transitivity of the causal relation), not merely a state with "some content or other." Similarly, my turning on the washing machine is causally sufficient for the specific way the clothes get jostled about. Thus, it is false that we lack control in the sense of *v*'s causal insufficiency for the outcome.

(3) *v does not reflect precise knowledge of r*. In turning on the washing machine, I may well know that the clothes will be jostled, but probably not the specific way. Similarly, for processes like calculating, problem solving, deliberation, and associative thinking, my volition to reach a content outcome reflects knowledge of the *type* of outcome I want, insofar as the volition represents that type of outcome. It does not reflect knowledge of the token outcome. For example, I know that I want to think *the integer equivalent to 288/12* or *the fastest way to the post office* or *the kind of veneer I like best for the desk* in willing, respectively, to think the thoughts whose contents satisfy these descriptions. I do not know that I want to think *24*, *Main Street express bus*, and *teak*. This sense of "lack of intentional control" is expressed in the passage quoted in my discussion of interpretation (2): "I'm going to think something — I don't yet know what" — that is, a lack of epistemic access to the ultimate result of catalysis. Now, "not knowing what you're doing" is, in a sense, lack of control. But we generally *do* know what we're going to do (what we're going to think) in directed cognition, even if not precisely. Indeed, to know the outcome precisely would obviate the need for the cognitive process. Thus, full epistemic access to the outcome is an unreasonable constraint on intentional control.

(4) *v is merely a distal cause of r*. "One can make such an event [a thought or imagining] occur, but only by doing something else" (p. 239). On this understanding, *v* lacks intentional control over *r* in the sense that it does not *directly*

cause *r*; it merely initiates the particular mechanism that delivers *r*. On the other hand, *v* itself is under intentional control insofar as it is done in a nonmediated way. This claim certainly seems to have introspective support: we do not exercise our will "by doing something else"; we simply will. And this is indeed a kind of control that we *lack* over the products of ballistic processes. Neither does it seem that we bring about the catalytic events that follow volition (e.g., the "inclining of the mind" following the volition to solve a problem) in a mediated way. On this understanding of intentional control, only *v* and *c* count as intentionally controlled. Thus Strawson writes, "the content outcomes are delivered into consciousness so as to be available in their turn for use by *the catalytic machinery that is under intentional control*" (p. 234; emphasis added).

Thinking "By Doing Something Else"

Evidently (4) is Strawson's (2003) main metaphysical argument for NATT, as he offers many parallels with physical behavior to support it. These are cases where the agent does something physical "by doing something else":

> But the event of entertaining itself is not an action, any more than falling is once one has jumped off a wall. (p. 235)

> Your thinking that *a* is *G* can be allowed to be the product of an action or actions performed with the intention to produce that particular thought-content, but it is not itself an action, any more than an increase in one's physical fitness is when one goes in for regular exercise. (p. 236)

> To think that the actual content-issuing and content-entertaining that are the heart of imagining are themselves a matter of action seems like thinking, when one has thrown a dart, that the dart's entering the dartboard is itself an action. (pp. 242–243)

> There is no direct action in the actual issuing of new content, any more than there is in the growth of trees one has planted. (p. 243)

In these examples, falling, increasing fitness, the dart's entering the dartboard, and the growth of trees are clearly all things one causes indirectly. The mediating processes operate ballistically once they are started; e.g., once the tree is planted, it begins to grow without further intervention from the agent. Suppose that one intends to fall, get fitter, stick the dart, and grow the tree, respectively, and that these volitions distally cause their effects. Are the effects not things one *does*, simply because they are accomplished via a ballistic process? A slippery slope threatens if we disqualify them as actions on such grounds. For even the intentional raising of one's arm — a paradigm physical action — is accomplished via a neuromuscular process that is ballistic, as Strawson himself notes: "Much bodily movement is ballistic, relative to the initiating impulse; the same goes for

thought" (2003, p. 245). This greatly restricts what qualifies as an action. It may well follow that *no* kinds of mental or physical events are actions, with the exception of volition and catalysis.

Given the connection between agency and responsibility, theories that minimize agency should not be adopted without careful consideration of less extreme alternatives. For example, we can instead distinguish between actions the agent does directly and actions she does via ballistic processes. As Strawson suggests, "perhaps the only error that some people make, in considering these matters, is to conceive of the issuing of a particular thought-content as a 'basic' action: something one does, and does intentionally, and does not do by doing anything else" (p. 236). But this implies that the content outcomes of ballistic processes should be conceived as "non-basic actions" — actions nonetheless.

Furthermore, the claim that the agent brings about volition and catalysis in a nonmediated way is itself problematic. No doubt they *appear* to be under nonmediated control; that is, there is no subjective phenomenon of producing one's volitions and catalytic states *via other mental states or processes*. Perhaps they do not even feel caused by us directly; that is, they feel simply uncaused. As Rosenthal (2002) maintains, "We are seldom if ever conscious of the mental causes of our conscious volitions. And that results in those volitions seeming spontaneous and uncaused" (p. 219). But surely they are in fact proximally and distally caused by beliefs and desires. For example, my volition to figure out the fastest way to the post office is proximally caused by my desire to save time and my desire to get to the post office, and distally caused by my desire to mail a package, my belief that there is a post office in the local area, that post offices mail packages, etc. Thus, volitions to think thoughts, like their content outcomes, are controlled via a series of mental states and the "natural causality of reason," to use Strawson's phrase. Moreover, suppose there are neural antecedents that causally determine conscious volitions to perform mental acts, similar to the "readiness potential" occurring milliseconds prior to conscious volitions to move (Libet, 1983). These neural precursors, which Strawson himself acknowledges (see p. 245), surely comprise a ballistic process insofar as they are subpersonal. Our conscious volitions, despite introspective appearances, would then be performed "by doing something else," namely, the initiation of that neural process.

But even if volition and catalysis *were* under nonmediated control and their outcomes lacked intentional control in that sense, it is unclear why mediated control would be a *less agentive* process, as Strawson appears to contend. Both varieties of control would ensure that cognition operates effectively, and neither exclusively involves subpersonal neural processes, which are regarded as not being constitutive of the agent. So, even if there were such a thing as nonmediated control over volitions, there is no reason to imply that such control is a higher level of mental agency.

Conclusion

I have argued against both Strawson's phenomenological and metaphysical arguments for the non-agentive thought thesis (NATT). The phenomenological argument is undercut by the empirical possibility of cognitive penetration; namely, Strawson's passivist phenomenology may be the product of his believing (i) and (ii), as well as his commitment to (iii). If that cognitive penetration obtains, the phenomenology would not support NATT independently of those beliefs — even if those beliefs were plausible.

Strawson's metaphysical grounds for NATT is based on (iii), specifically the claim that we bring about content outcomes merely in an indirect way. I have argued that requiring direct control of any mental event that is to count as an action is an unreasonable constraint: such control arguably does not exist, and if it does, only volition and catalysis would seem to qualify as actions — excluding all the mental and bodily behavior they cause. There is no principled reason to so greatly restrict the scope of action, when instead we can simply distinguish between basic and non-basic (mediated) action.

In sum, the debate about mental agency is best pursued in the metaphysical arena. Inferences from phenomenology to the metaphysics of mind are notoriously unreliable, particularly under Physicalism. To give one example, the phenomenological unity of conscious experience may lead us to expect to find a single locus of consciousness in the brain, a "Cartesian theater" that is the substrate of all conscious representations at a given time. But there is no empirical support for such a locus, and the idea is at odds with what we know about neural information processing, as Dennett (1991) argues. A second example, noted above, is the experience of our own volitions as being uncaused, which may lead us to the arguably mistaken idea that they are in fact uncaused (Rosenthal, 2002) or caused in a direct way by the agent. A further problem for the phenomenological approach is simply that phenomenologies may support clashing metaphysical views. For example, to the question — does thinking feel like willful activity? — "answers are surprisingly varied," Proust notes (2009, p. 253). In that case, we arrive at a theoretical impasse, as no phenomenology is inherently "wrong"; a given phenomenology may be the result of "careless" introspection, but how can we establish this? Lastly, a phenomenology that supports a certain metaphysical claim may be penetrated by the subject's belief in that claim, as I have attempted to illustrate with regard to Strawson's passivist phenomenology.

Of course, the possibility of cognitive penetration undermining a phenomenological argument for a metaphysical claim is not restricted to Strawson's case. Nor do I intend to imply that only beliefs (i)–(iii) can penetrate the sense of mental agency; perhaps the anti-free will beliefs instilled by Lynn et al. (2014) can also reduce the sense of mental agency, for example. These possibilities invite the question, is there a "natural," i.e., non-penetrated phenomenology of mental agency?

Supposing the subject holds *no* beliefs on the nature of mental agency that could affect the character of her phenomenology, I would hypothesize she would feel some degree of agency for all cases of (successful) directed cognition, based on a perceived match between volitional content and the cognitive outcome, and a match in a subpersonal comparator (if there is one for mental acts). But even the nonpenetrated case — the "default" setting for felt agency, as it were — would not be the proper starting point for an inquiry into which mental events, if any, are actions. I maintain that the inquiry should begin with the metaphysics of mind, specifically by considering the relationship between volitions to think, ballistic processes, and content outcomes, as well as how the self is related to the processes that yield those outcomes.

References

Adams, F. (1997). Cognitive trying. In G. Holmstrom–Hintikka and R. Tuomela (Eds.), *Contemporary action theory* (Vol. 1, pp. 287–314). Dordrecht, The Netherlands: Kluwer.

Buckareff, A. A. (2005). How (not) to think about mental action. *Philosophical Explorations, 8*(1), 83–89.

Campbell, J. (1999). Schizophrenia, the space of reasons, and thinking as a motor process. *The Monist, 82*, 609–625.

Dennett, D. (1991). *Consciousness explained*. Boston, Massachusetts: Little Brown.

Desantis, A., Roussel, C., and Waszak, F. (2011). On the influence of causal beliefs on the feeling of agency. *Consciousness and Cognition, 20*, 1211–1220.

Dorsch, F. (2009). Judging and the scope of mental agency. In L. O'Brien and M. Soteriou (Eds.), *Mental actions* (pp. 38–71). New York: Oxford University Press.

Farrer, C., Franck, N., Georgieff, N., Frith, C .D., Decety, J., and Jeannerod, M. (2003). Modulating the experience of agency: A positron emission topography study. *NeuroImage, 18*, 324–333.

Feinberg, I. (1978). Efference copy and corollary discharge: Implications for thinking and its disorders. *Schizophrenia Bulletin, 4*, 636–640.

Frith, C. D., Blakemore, S.-J., and Wolpert, D. M. (2000). Abnormalities in the awareness and control of action. *Philosophical Transactions of the Royal Society of London Series B – Biological Sciences, 355*, 1771–1788.

Haggard, P., Clark, S., and Kalogeras, J. (2002). Voluntary action and conscious awareness. *Nature Neuroscience, 5*, 382–385.

Kim, J. (1989). Mechanism, purpose, and explanatory exclusion. *Philosophical Perspectives, 3*, 77–108.

Libet, B. (1983). Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential). The unconscious initiation of a freely voluntary act. *Brain, 106*, 623–642.

Lynn, M. T., Muhle–Karbe, P. S., Aarts, A., and Brass, M. (2014). Priming determinist beliefs diminishes implicit (but not explicit) components of self-agency. *Frontiers in Psychology, 5*, 1483, 1–6.

Marcel, A. (2003). The sense of agency: Awareness and ownership of action. In J. Roessler and N. Eilan (Eds.), *Agency and self-awareness* (pp. 48–93). Oxford: Oxford University Press.

Mele, A. (2009). Mental action: A case study. In L. O'Brien and M. Soteriou (Eds.), *Mental agency* (pp. 17–37). New York: Oxford University Press.

Metzinger, T. (2004). Conscious volition and mental representation: Toward a more fine-grained analysis. In N. Sebanz and W. Prinz (Eds.), *Disorders of volition* (pp. 19–48). Cambridge, Massachusetts: MIT Press.

Pacherie, E. (2000). The content of intentions. *Mind & Language, 15*(4), 400–432.

Pacherie, E. (2007). The sense of control and the sense of agency. *Psyche, 13*(1), 1–30.

Peacocke, C. (2007). Mental action and self-awareness (I). In B. McLaughlin and J. Cohen (Eds.), *Contemporary debates in philosophy of mind* (pp. 358–376). Oxford: Blackwell.

Proust, J. (2009). Is there a sense of agency for thought? In L. O'Brien and M. Soteriou (Eds.), *Mental agency* (pp. 253–279). New York: Oxford University Press.

Rosenthal, D. M. (2002). The timing of conscious states. *Consciousness and Cognition, 11*, 215–220.
Schlosser, M. E. (2012). Causally efficacious intentions and the sense of agency: In defense of real mental causation. *Journal of Theoretical and Philosophical Psychology, 32*(3), 135–160.
Seli, G. (2012). The utility of conscious thinking on higher-order theory. *Philosophical Explorations, 15*(3), 303–316.
Siegel, S. (2012). Cognitive penetrability and perceptual justification. *Nous, 46*(2), 201–222.
Strawson, G. (1997). "The self." *Journal of Consciousness Studies, 4*(5–6), 405–428.
Strawson, G. (2003). Mental ballistics *or* the involuntariness of spontaneity. *Proceedings of the Aristotelian Society, 103*(3), 227–256.
Strawson, G. (2009). *Selves: An essay in revisionary metaphysics*. Oxford: Oxford University Press.
Synofzik, M., Vosgerau, G., and Newen, A. (2008). Beyond the comparator model: A multifactorial two-step account of agency. *Consciousness and Cognition*, *17*, 219–239.
Wegner, D. M. (2002). *The illusion of conscious will*. Cambridge, Massachusetts: MIT Press.
Wegner, D. M., and Wheatley, T. (1999). Apparent mental causation: Sources of the experience of will. *American Psychologist, 54,* 480–491.
Yablo, S. (1992). Mental causation. *The Philosophical Review, 102*(2), 245–380.