Review of Philosophy and Psychology (2020) 11:157–181
https://doi.org/10.1007/s13164-019-00449-6

A Modified Self-Knowledge Model of Thought Insertion

Sruthi Rothenfluch[1]

Published online: 30 October 2019

Abstract

Thought insertion is a condition characterized by the impression that one's thoughts are not one's own and have been inserted by others. Some have explained the condition as resulting, in part, from impaired or defective self-knowledge, or knowledge of one's mental states. I argue that such models do not shed light on the most puzzling feature of thought insertion: the patient's experience that an introspected thought does not feel like her own. After examining ways in which existing versions of the model might address this worry, I propose a significant modification. I argue that the experience of disownership consists in a rational indifference that one feels towards one's inserted thought. I further contend that the experience is generated by an underlying absence of an expectation of rational authority towards the inserted thought, such that the patient does not expect her thought to reflect, or be shaped by, her own rational considerations. I defend my proposal using empirical studies from cognitive and social psychology which suggest that we ordinarily have and experience an expectation of rational authority towards a certain subset of our thoughts, and direct analysis of patient reports, which strongly suggest that it is this expectation and the corresponding experience that thought insertion patients lack.

Thought insertion is a condition primarily associated with schizophrenia in which patients have the impression that some of their thoughts are not their own and have been inserted by others.[1] Such patient reports are difficult to fully grasp because patients appear to have first-person access to a thought (a mode reserved for the subject of the thought) that they claim does not feel like their own. In other words, gaining

[1]Thought insertion is commonly associated with schizophrenia in philosophical, psychological and neurological discussions of the condition: Mellor, 1970; Frith, 2003; Sims, 2003; Fernandez, 2010.

✉ Sruthi Rothenfluch
rothenfl@up.edu

[1] Department of Philosophy, University of Portland, 5000 N Willamette Blvd, Portland, OR 97229, USA

awareness of a thought through direct introspection would be thought to automatically produce the impression that it is one's own.[2] How, then can we make sense of this report? How can a subject be first-personally aware of a thought that she does not experience as her own?[3]

Some have explained the condition as resulting, in part, from impaired or defective self-knowledge, or knowledge of one's mental states. While such models reflect a distinctive feature of thought insertion, namely, a radical detachment towards one's inserted thought, they do not shed light on the most puzzling feature of thought insertion: the patient's experience that an introspected thought does not feel like her own. What we need is an account that tells us not only *why* patients experience a thought as not their own but an explanation of what *constitutes* this experience. Further, this experiential analysis must distinguish the experience of thought insertion from more mundane episodes in which one might feel distanced or disconnected from their self-attributed thought. I argue that the self-knowledge model must be significantly modified in order to satisfy both desiderata. In particular, I argue that the experience of dis-ownership, or the impression that a thought is not one's own, consists in a rational indifference that one feels towards one's inserted thought. Crucially, what such rational indifference feels like for the thought-insertion patient, and what the ordinary sense of mine-ness feels like for the non-afflicted, will depend on other features of the situation. As such, my account involves a relatively fine-grained approach to understanding the experience of mine-ness and its loss in thought insertion. I further contend that the experience is generated by an underlying absence of an expectation of rational authority towards the inserted thought, such that the patient does not expect her thought to reflect, or be shaped by, her own rational considerations. This model elucidates the *experience* of disowning one's thought while also distinguishing the condition from more ordinary, day-to-day forms of alienation, or so I will argue.

I will begin in the first section by giving the reader a sense of the condition by presenting and directly analyzing patient reports. As my discussion will not encompass all the features of thought insertion, I will also in this section clarify and defend my specific explanatory objective, namely, to provide an analysis of the experience of dis-ownership. Next, I will present an overview of two prominent self knowledge (SK) models of thought insertion and argue that such models are explanatorily inadequate, and proceed to explore means by which these accounts might by-pass my objection. I argue, however, that such attempts prove unsuccessful because they fail to distinguish the experience of dis-ownership, characteristic of thought insertion, from the more mundane experience of having irrational thoughts. In this section, I will introduce a third self-knowledge model that appears to avoid the criticisms leveled against the other two, but ultimately argue that this model also fails to provide an adequate explanation of the experience of dis-ownership. Then, I will present and defend my own positive proposal, according to which the experience of thought insertion consists in a felt rational indifference towards one's thoughts that is itself caused by an absence of the expectation of rational authority. Finally, I will address a number of legitimate and

[2] Parrot, 2017, 42; Bortolotti and Broome, 2009, 214; Sollberger, 2014, 591.

[3] Some have cast the puzzle in terms of a potential violation of Shoemaker's immunity to error through misidentification principle, according to which one cannot be mistaken about the subject of a thought in making a self-attribution whenever 'I' is used as subject. That is, the person making the attribution might be wrong about the truth of the thought, but cannot be wrong that it is she who is having the thought.

potentially serious concerns against my view including the plausibility of an *ordinary* expectation of rational authority and whether or not it is the failure of such an expectation that underlies the experience of dis-ownership in thought insertion.

1 Understanding Thought Insertion

As stated above, thought insertion patients report that certain conscious thoughts are not their own, and instead have been inserted by a foreign agent:

> A 29-year-old housewife said, 'I look out of the window and I think the garden looks nice and the grass looks cool, but the thoughts of Eamonn Andrews comes into my mind. There are no other thoughts there, only his... He treats my mind like a screen and flashes his thoughts on to it like you flash a picture'. (Mellor, 1970, 17)
>
> Things are put into my mind, like 'Kill God'. It is just like my mind working, but it isn't. They are not my thoughts. They belong to this guy, Chris. They are his thoughts. (Frith 1992)
>
> All of a sudden these strange thoughts came into my head as though I should… I thought to kill the cat. I shrugged it away and walked off. Thoughts like that came into my head. It wasn't me that was thinking them, it was as though it was put there. When it first started I thought it was some kind of force of some sort. (Allison-Bolger, 2015, 237)

We can distinguish two aspects of the condition indicated in these reports: first, the patient describes having a thought that does not feel like his/her own, which I will refer to interchangeably as the experience of dis-ownership or the loss of mine-ness. Second, patients report the impression that the thought was inserted by someone else. Notice here that it is not merely that one's thoughts have been heavily *influenced* by another person. Where we might ordinarily say "my mother gave me the idea", the thought-insertion patient "believes that by some concrete process the boundaries of his self involving thinking are so invaded that his mother is actually placing thoughts inside of him" (Sims, 2003, 157).

There are two methodological points to be made at the outset. First, my analysis will focus specifically on the first component. I will defend an account that both explains what this experience consists in and how it is produced. While my model may be compatible with various accounts of misattribution (ascribing the thought to a foreign agent), I will not concentrate on this aspect. Why limit my discussion in this way? Outside of the dialectical point that the accounts I target attend exclusively to the experience of disowning one's thought,[4] patients, as evidenced by various reports cited throughout this discussion, often report a loss of mine-ness, without misattribution. What is consistent across patient reports, and arguably essential to the condition, then, is the loss of mine-ness. Furthermore, it is this aspect of thought insertion that seems most philosophically puzzling. This is because accessing one's thought within one's stream of consciousness would

[4] Pickard 2010; Bortolotti and Broome, 2009; Fernandez 2010

generally be accompanied by the implicit impression that the thought is one's own. It is this "divorce between first-personal awareness of the content of a thought, and the possibility of self-ascribing that thought" that makes it difficult to have a firm understanding of what the patient is experiencing. (Bortolotti and Broome, 2009, 214). The problem is exacerbated by the fact that we cannot subsume the condition under other, easier-to-imagine forms of detachment. For example, in non-pathological cases, subjects may feel distanced from their conscious thoughts because they are ashamed, disappointed or frustrated by them. We can imagine and relate to such experiences. But it is not this kind of disapproval-induced alienation or detachment that thought-insertion patients report. Patients report that some of their thoughts simply (and perhaps non-figuratively) do not feel like their own. In cases of pathological detachment outside of thought insertion, we might be able to imagine what it is like to hear voices to some extent, by imagining hearing sounds or speech that we are not responsible for. Thought-insertion patients, however, deny that they hear voices (Parrot, 2017, 42; Saks, 2007, 84). Thus, the condition cannot be characterized or understood in these terms.

Second, my objective is to examine a particular approach to thought insertion and examine its prospects for addressing what appear to be significant short-falls. I will ultimately argue that the self-knowledge model can be modified and fruitfully developed to yield an illuminating portrayal of the experience of dis-ownership. As such I will not be offering a comparative survey of various models and adjudicating between these.[5]

2 Thought Insertion as an Impairment of Self-Knowledge

Bortolotti and Broome (2009) and Pickard (2010) have argued that thought insertion reflects a failure of self-knowledge, or knowledge of one's own mental states. They draw from Richard Moran's (2001) conception of ordinary self-knowledge according to which one has proper knowledge of one's judgment-sensitive attitudes (exclusive of sensory experiences and 'brute' desires such as those associated with hunger and thirst) to the extent that one exercises rational authority over these attitudes.[6] Having rational authority amounts to the capacity to rationally determine, or make up one's own mind about, what one believes, fears, desires, intends, etc. Having rational authority does not mean that one can at *will* form their desires or beliefs, nor does it mean that every attitude must be a product of explicit rational deliberation. Rather, Moran clarifies that a subject acquires ordinary self-

[5] John Campbell (1999) takes a similar approach with respect to Frith's comparator model; Graham and Stephens (1994) exclusively discuss and defend the sense of agency model; Sollberger (2014) explains and defends the endorsement model of thought insertion from objections.

[6] Judgment-sensitive attitudes are those that are sensitive to a particular type of reason, those for which "reasons in the standard normative sense can sensibly be asked for and offered" (Scanlon, 20). When I am asked why I have a backache, I might be able to cite a bad car accident or bad posture, but these will amount to causal explanations and "do not seek to say anything about the apparent point, or the good, or the intelligibility of the state I am in" (Moran, 2012, 214). In contrast, when we ask *why* of judgment-sensitive attitudes, we are asking for the reasons for one's beliefs or what is good or worthwhile about the objects of one's desires. It is not only that it is appropriate to ask why in the sense indicated, but that the answer to this question also tells us how the person came to have that attitude.

knowledge when he takes his attitudes to be "answerable to, and adjustable in, the light of his" reasons (118). Proper knowledge that one has a desire, for example, entails that one is "not only prepared to justify this desire if asked, but the presence or absence of justification makes a difference to the presence or absence of the desire itself, and the direction of [one's] desire is in fact guided by the direction of [one's] thought about what is desirable" (118). Thus, according to Moran, a subject properly knows that she has a particular attitude when she views that attitude as determined and shaped by her own normative assessments. In contrast, acknowledging the presence of an attitude that is impervious to one's own reasons is an improper or impaired form of self-knowledge, where one neither experiences commitment to, nor responsibility for, that state.

How is this conception put to use in the analysis of thought insertion? Bortolotti and Broome (2009) argue that patients with thought insertion have impaired knowledge of their inserted thoughts. This impairment is manifested by the patient in two ways: the failure of ownership and the failure of authorship. The failure of *ownership* is defined in terms of an inability to self-ascribe a thought that has been introspectively accessed; the patient "doesn't acknowledge it as hers," and views her inserted thought as "unwanted and unwelcome" (219). Bortolotti and Broome characterize a lack of *authorship* as an inability to endorse the content of one's thought, which is intended to mirror Moran's conception of rational authority. Having authorship of one's thought does not require that one formed the thought through active deliberation, but rather that "one be able to endorse that belief on the basis of the best evidence available to him" (212).[7] In this way, the subject takes responsibility for the thought and makes a commitment to it. According to Bortolotti and Broome, the thought-insertion patient lacks this capacity as she "cannot give reasons for endorsing the content of her inserted thought" (222).

Pickard (2010) similarly explains thought insertion as a form of impaired self-knowledge. She maintains that the experience that a thought is not one's own is caused by the fact that such thoughts are manifestations of states that one does not rationally support: "looking outwards to the world, they judge that the mental states… are not warranted or appropriate: they do not reflect how the world actually is or should be" (67). While we ordinarily dismiss or revise such states, thought-insertion patients are struck by a "radical persisting disparity" between the inserted thoughts that appear within their consciousness and states to which they are rationally committed, which "is why they disown these manifestations—why they do not believe that these … are caused by their own mental states" (67-8).

According to SK models, the patient is aware of the inserted thought, but such awareness falls short of ordinary self knowledge either because the patient does not endorse its content (Pickard), or cannot rationally endorse and cannot self ascribe the thought (Bortolotti and Broome). These models are, however, explanatorily deficient in that they leave unanalyzed the *experience* of dis-ownership. While Bortolotti and Broome accept that patients exhibit a failure of ownership towards their inserted thought, define this failure as an inability to self-ascribe the thought, and claim that "it is mineness which is conspicuously missing from the subject's phenomenology" (216), they do not tell us what this loss of mine-ness consists in. The authors essentially

[7] Bortolotti and Broome maintain that authorship is necessary for ordinary self-knowledge only when it comes to attitudes that are central and meaningful to one's life.

take patients' experiential descriptions at face value. While this is not in and of itself bad practice, in this case doing so is uninformative and unhelpful because the report does not straightforwardly correspond to an understandable experience. Part of what is mysterious about thought insertion is precisely what the experience itself amounts to: what does it feel like to access a thought within one's own stream of consciousness that does not appear to be one's own?

Pickard, likewise, treats the report of dis-ownership as self-explanatory and moves on to provide a causal explanation of the experience. She argues that it is a perceived disparity between one's inserted thoughts and one's subjectively justified attitudes that produces the experience of dis-ownership, but does not tell us what the experience itself consists in. Pickard contends, however, that this experience is not as elusive as I've suggested, arguing that we commonly have experiences that lie on the same continuum. We "reach for the idea of failures of ownership and identification" and "often wish to dissociate ourselves from "immoral, selfish or shameful thoughts (59). Such remarks seem to both mischaracterize ordinary experiences and fail to appreciate the pathological nature of the delusion. In the ordinary case what is troubling is not that these thoughts are *not* one's own, but the realization that they are. That is, I might *wish* that certain thoughts would not occur to me; I might feel disappointed or ashamed that I am capable of having certain thoughts of which I disapprove.[8] But such second-order reactions betray my sense that these first-order thoughts are, regrettably, mine. Such situations are starkly different from the experience of dis-ownership in which patients do not merely view their thought as unwanted or distasteful —a non-pathological and relatively common reaction to one's thoughts—but as not their own.[9] It is this exceptional, pathological loss of mine-ness that is left unexplained by Pickard.

3 Potential Solutions

3.1 Re-Thinking the Role of Rational Authority

One might wonder whether the fix for SK models is to simply utilize Moran's conception differently within their account. Rather than identifying an absence of rational authority as a causal explanation of the experience of dis-ownership—as Pickard does—or as a symptom that is independent of the failure of ownership—as Bortolotti and Broome do, SK theorists might instead identify the absence of rational authority (an inability to endorse one's inserted thought) as constitutive of the *experience* of dis-ownership. The idea here would be that the patient's experience that a

[8] This is also true of other mental illnesses in which patients feel disconnected from their thoughts, but do not disown them. Billon (2013) maintains that OCD patients find themselves troubled by the fact that their recurring, unwanted thoughts might in fact reflect their true character: "It is because the patient (rightly) self-attributes the intrusive thought that he fears it might be representative of his personality" (297). Churchland also points out that although the smoker "might wish that [the desire for another cigarette] were not his, but so far as the feeling itself is concerned, it is as much his as his desire to quit smoking"(2002, 209). In both cases, patients disapprove of their thoughts and *wish* they were not their own, but do not disown them. Their distress arises from the fact that such thoughts are genuinely representative of their character.

[9] In fact, as will be shown shortly, some patients endorse the content of their inserted thoughts.

thought is not her own just is the sense that she cannot rationally endorse the content of her thought.

Before evaluating this move, it is important here to clarify the difference between it and the original accounts. Bortolotti and Broome, recall, accept that thought-insertion patients exhibit a failure of ownership which they define as an inability to ascribe the inserted thought to oneself or view it as *mine*. However, on their view, the inability to ascribe the thought to oneself is not the same as, nor is it correlated with, the failure of authorship, defined as the inability to endorse one's thought: "when granting the capacity for self ascription, no additional claims need to be made about the thought being endorsed by the subject" (216). Thus, on Bortolotti and Broome's original account, the loss of mine-ness does not consist in the experience of being unable to rationally justify one's thought; rather, these are independent elements of thought insertion. According to Pickard, the inability to endorse one's inserted thought is responsible for producing the sense that the inserted thought is not one's own, but the former is not offered as a constitutive explanation of the latter. The proposed modification is that the experience that a thought is not one's own consists in the impression that one cannot rationally justify one's thought.

While amending the SK model in this way offers a potential explanation of the experience of dis-ownership, it faces other serious problems. First, the modified account implies that the inserted thought will invariably clash with the patient's own normative considerations, but this is not in fact borne out by patient reports:

> [He] said […] 'it's like a thought as it comes in…a thought is very light really... it's a light feeling where you feel as though I'm actually thinking it…it's just a thought but it feels logical say… it feels pretty normal or fits with what I suspect…I wonder if that's me…it felt like a piece of information…I don't think that was mine (Hoerl, 2001)
>
> [She] said that sometimes it seemed to be her own thought 'but I don't get the feeling that it is'. She said her 'own thoughts might say the same thing', but the feeling isn't the same', the feeling is that it is somebody else's'. She asked if she had other people's thoughts put inside her head. She said 'possibly they are but I don't think of them in that way…. They were being put into me into my mind….very similar to what I would be like normally. (Hoerl, 2001)

These patients do not appear to object to, or disagree with, the content of their inserted thoughts. Instead, they find their respective thoughts consistent their own deliberations, one claiming that his thought 'feels logical' and 'fits with what [he] suspects' and the other noting that her thought is "very similar to what [she] would be like normally'. Both patients nevertheless report that such thoughts do not feel like their own. Thus, the experience of dis-ownership does not seem to boil down to an inability to endorse one's thought.

Second, it is not uncommon to believe that one has a thought whose content one does not endorse. For example, one might acknowledge that one has a fear of flying even when one knows that the probability of accidents during flight is significantly lower than the probability of car accidents. Outside of irrational phobias, a subject might self-ascribe an irrational attitude because she finds herself so utterly committed to it in her practical, emotional, and intellectual life

that she cannot bring herself to change the attitude. In such cases, one "continues to believe something (continues to regard it with conviction and to take it as a premise in subsequent reasoning) even though he or she judges there to be good reason for rejecting it (Scanlon, 1998, 25). Suppose that Karen believes that she has an aptitude for science and music, but not for history and philosophy, based on the results of an aptitude test.[10] Karen plans her life around this belief and develops a firm feeling of conviction in the truth of this belief. However, she later discovers that she received someone else's test scores. Karen ought to give up her belief, but it seems likely that Karen would continue to believe that she had a great aptitude for science and music but not for philosophy and history. It is not that Karen does not know about the scores, or that she is unaware of her belief. Rather, she both understands that she lacks justification for her belief and acknowledges that she nevertheless maintains the belief. This suggests that the experience that a thought is not one's own cannot be the same as an inability to justify the thought. We need not rely on our intuitions in hypothetical cases to see this: incoming students at my religiously-affiliated university often grapple with their belief in God's existence after being introduced to powerful arguments contesting his existence. In some cases, students continue to self-attribute belief in God, despite acknowledging the probative force of these counter-arguments. These students, in other words, view certain beliefs as their own despite an inability to justify tem. Therefore, the experience of dis-ownership, the sense that a thought is not mine, cannot consist in an inability to justify one's thought.[11]

3.2 Fernandez's Commitment and by-Pass Model

Let us take stock. I've argued that original SK models do not provide an explanation of the experience of dis-ownership and that characterizing the experience in terms of an inability to endorse the inserted thought will not help, as some patients view their inserted thoughts as justified. In addition, we can easily call to mind actual and hypothetical cases in which subjects regard certain beliefs as their own despite an inability to justify them. Before I move on to my own proposal, I want to consider a different self knowledge model that offers an explanation of the experience of dis-ownership and its causal source.

According to Fernandez (2010, 2013), when we attribute attitudes to ourselves under ordinary circumstances, we feel committed to their content. For example, "if I determine … that one of my beliefs is that my wife is cheating on me, then that belief is not presented to me as being neutral on whether she is actually cheating on me or not. That belief is presented to me as being correct" (Fernandez, 2013, 167). Further, the reason that we have this experience is that we base our second-order attitudes (our beliefs about what attitudes we have) on our grounds for our first-order attitudes, a method he calls by-

[10] Cassam, 2014.

[11] In fact, the problem of recalcitrant attitudes undermines not only my modified version of the SK model, but also Pickard's original claim that the experience of dis-ownership is produced by the disparity between one's conscious thoughts and one's considered judgments, as this does not seem to occur in the case of recalcitrant attitudes.

pass. For example, suppose I believe that I believe that climate change is having a disproportional effect on developing nations. My grounds for my second-order belief is the same as my grounds for my belief that climate change is having a disproportional effect on developing nations. For this reason, I am immediately compelled to endorse the content of my self-attributed belief. Once I ascribe the belief to myself, it is no longer an open question for me whether or not climate change is having a disproportional effect on developing nations. Fernandez maintains that thought insertion patients experience an absence of this commitment in that they "do not experience those beliefs as forcing any particular picture of the world" on them (Fernandez, 2010, 78). Fernandez argues that the reason that thought-insertion patients do not experience commitment is because they do not form their second-order beliefs using by-pass (84).

Fernandez's model is preferable to the other SK models on many counts: first, he offers an explanation of the experience of dis-ownership as an absence of the ordinary commitment we feel towards our self-attributed beliefs. Second, he offers a causal explanation of how patients come to have such an experience. Third, unlike the modified SK model, Fernandez's account can accommodate reports in which patients appear to endorse the content of their inserted thoughts. Fernandez argues that such patients observe a coincidental overlap between their perception of the world and their inserted thoughts, but they do not base their awareness of their inserted thought on the grounds for the thought.[12] Subsequently, their awareness of their inserted thought itself does not compel them to endorse the content of this thought, as evidenced by locutions that express only a "highly contingent" and "loose" relationship between their beliefs and their inserted thoughts.

Recalcitrant attitudes, however, remain a problem for Fernandez's account. To see why, let us take a closer and more nuanced look at the complex situation in which the young religious student finds herself. The student believes in God and knows that she does so despite acknowledging the force of arguments against God's existence. But we might imagine that such arguments give rise to some disquiet. Perhaps certain personal experiences or occasional discrepancies within religious teachings redirect her attention to the counter-arguments, bringing her once latent doubts to salience. There is a sense in which the subject is not *wholly* convinced of the truth of her belief. We can imagine the same thing occurring in Karen's situation: when she makes an error from time to time within the context of science and music, she might casually question her aptitude, wondering whether her actual test results showed aptitude in other subjects. Commitment, then, does not appear to be all or nothing. In these cases, the subject believes that *p* for the most part—Karen plans her career and uses her science-and-music-aptitude belief in much of her reasoning; the student identifies herself as a believer and structures her life and values accordingly. However, characterizing the two as wholly committed to their beliefs is to oversimplify their stance. There are times in their lives where the relevant beliefs are seriously questioned.

[12] He likens such access to a third-person perspective of the type that a psychotherapy patient might have to her mental states.

These periods of doubt and dissonance may be brief and eventually ignored, but they nevertheless suggest that such subjects are not thoroughly committed to the truth of their beliefs.

What does this say about Fernandez's model? The subjects in these cases view these beliefs as their own: Karen takes herself to believe that she as an aptitude in science and music and the student believes that she believes in God. Despite this, they do not feel wholeheartedly committed to the truth of their beliefs, manifested by their tendency to periodically attend to their own misgivings. These doubts are not strong enough, in the end, to cause them to abandon their belief—perhaps because they have become so deeply interwoven in their lives—but they are there. This suggests that the *experience* of ownership does not consist in the feeling of commitment that one has towards the content of one's states, as these subjects view their beliefs as their own even though they are not compelled to fully endorse their content.[13]

4 An Alternative: A Modified Self-Knowledge Account

I want to propose a significant modification to the SK model which will both provide an explanation of the experience of dis-ownership and avoid the problems generated by recalcitrant attitudes. Ordinarily, we expect to have rational authority over our judgment-sensitive thoughts. What this means is that we expect (though this expectation might not always be met) our beliefs, desires, intentions, etc., to be rationally justified by our own lights, i.e., subjectively justified. This expectation of rational authority is experientially manifested in different ways, depending on whether or not the expectation is met. In the ideal situation, the expectation is experienced negatively, such that one perceives no distance or separation between one's normative assessments and the attitudes that one self-attributes. Suppose that I am evaluating a certain presidential candidate, wondering about the likelihood that this candidate will beat the incumbent. I might call to mind her charisma, bi-partisan support for her policy proposals, and her highly-successful grassroots fund-raising efforts. Once I have considered these aspects of her campaign in order to determine whether she stands a good chance of beating the incumbent, I will not then feel the need to determine what *I believe* about her chances of beating the incumbent. From my perspective, these deliberations go hand in hand. It is this seamlessness that captures the phenomenology of the satisfied expectation of rational authority. The experience of seamlessness is that of moving automatically from one's normative judgment (that p is true/desirable/to be feared) to self-attribution of the corresponding

[13] One might wonder whether subjects in such cases experience a weakened sense of ownership. That is, Fernandez might argue that one's experience of ownership tracks the degree of commitment to one's self-attributed belief. Thus, the subjects' diminished commitment to their belief maps onto their weakened sense of ownership. This view, however, presupposes that the experience of ownership itself comes in degrees, which seems implausible. To be clear, I may doubt the truth of a belief that I attribute to myself, which might lead to some cognitive tension. However, throughout the process, I will not begin to view the weakened belief as only *slightly* my own. The experience of *mineness* does not shift in this way.

attitude (I believe that *p*/ desire *p*/fear that *p*).[14][15] In these cases, the experience of mine-ness, that is, the sense that a thought is one's own, consists in the feeling of seamlessness.

The expectation of rational authority is experienced differently when it is unmet. I might, for example, knowingly maintain a belief that is unjustified, fear something that poses no actual danger, or desire what is harmful or disadvantageous. We do not acknowledge such discrepancies with neutrality and acceptance, but rather sense a distinctive discomfort or agitation at their presence—the degree of which will vary according to how important and impactful the attitude is in our lives. Such uneasiness might then direct us towards means of alleviating the tension, by say, restructuring our normative standards.[16] I might, for example, re-evaluate what I consider to be harmful in order to render my formerly irrational desire more acceptable, or lower the credibility of counter-evidence against my existing beliefs. Without doing so, the discomfort persists. Miranda Fricker's (2007) example of a "card-carrying feminist" illustrates this experience: we are to imagine a woman freed from sexist beliefs and "yet she is influenced by a stereotype of women as lacking the requisite authority for political office, so that she tends not to take the word of female political candidates as seriously as that of their male counterparts" (Fricker, 2007, 37). Fricker imagines that once the individual becomes aware of this discrepancy, she critically examines her attitudes and "asks herself why" she affords women political candidates lower degrees of credibility than their male counterparts and takes efforts to "limit the impact of prejudicial residue on her credibility judgments" (38). Here, the subject's awareness of dissonance produces a distinctive tension which eventually prompts her to take certain measures to alleviate the discomfort.

Where the expectation of rational authority is unmet, the subject does not experience seamlessness, but rather a distinct disturbance, a feeling that something is amiss. This is what is experienced in the case of recalcitrant attitudes. Here, the subject's sense that the irrational attitude is her own consists in the uneasiness of dissonance, which pulls her towards finding a way to reconcile disparate elements in order to ameliorate her discomfort.

[14] As such, this experience is related to the notion of transparency. Transparency, as it is used by Gareth Evans and Richard Moran denotes a mode of access to our attitudes: one has transparent access to her attitudes when one knows what one believes or desires by directing one's gaze to the intentional object. Transparency, then, is a means by which one knows what one's attitudes *are* by considering reasons in favor of its content (such as the desirability of the object of desire, evidence for the truth of one's belief). Here, I am using the notion of *seamlessness* to account for the experience that one's attitudes are *one's own* under certain conditions, namely, where the thought in question meets one's expectation of rational authority. One way in which subjects experience seamlessness is by having transparent access to the content of their attitudes, but this is not the only way, as one can also experience seamlessness when one cannot recall one's reasons (see note 15).

[15] To be clear, the experience of seamlessness does not require that one be able to rehearse the reasons for one's attitude. Rather, one can experience seamlessness even in cases where one recalls one's attitude, but not the reasons for the attitude. If the subject remembers that her belief *was* rationally justified for her, then she might well experience her self-attributed belief as automatically justified. Suppose, for example, that Jalen believes that he believes that the assassination of Archduke Ferdinand was one of the causes of World War I, but does not remember the reasons for his belief. Now, he might recall that the belief was rationally justified for him. In this case, he will still experience seamlessness between his self-ascription and the content of the attitude. That is, he will believe not only that he believes that Archduke Ferdinand's assassination led to World War I, but also that it is true that Archduke Ferdinand's assassination led to World War I.

[16] I will discuss this phenomenon in greater detail in my discussion of cognitive dissonance theory.

In thought insertion, patients do not experience the seamlessness generated by the satisfied expectation of rational authority nor do they experience the tension produced by an unmet expectation of rational authority. Instead, they are left with a sort of rational apathy toward their inserted thought. Such apathy is experienced differently depending on whether the inserted thought happens to align or conflict with their own normative assessments. When they encounter inserted thoughts that appear justified, they do not experience seamlessness. That is, they cannot move directly from their normative assessment to awareness of their inserted thought, nor does awareness of their inserted thought immediately strike them as rationally justified. Rather, after determining their own normative position with respect to an issue, it is a further question for them whether or not they have the inserted thought. On the other hand, when patients encounter thoughts that conflict with their own normative assessment, they do not experience psychological discomfort that pushes them towards resolving the discrepancy. It is this rational apathy that constitutes the experience of dis-ownership. That is, the experience that an introspectively accessed thought is not one's own consists in the feeling of rational indifference towards the thought. Patients have this experience because they lack an expectation of rational authority; they do not expect these thoughts, unlike their other judgment-sensitive attitudes, to be subjectively justified.

5 A Defense of the Rational Expectation View

I have argued that ordinarily, we expect our judgment-sensitive attitudes to be rationally justified, that is, conform to our own normative reasons. This expectation when met, is experienced negatively, as an absence of a gap between one's normative assessments and one's attitudes: in deliberating about a question, the subject senses no distance between her conclusion and her self-attribution; in recalling a past rationally-justified attitude, the subject immediately views this attitude as one and the same as her normative judgment. When the expectation is unmet, the subject experiences psychological discomfort that motivates them to adjust their cognitive structure in a way that alleviates the tension. These constitute the experience of mine-ness we feel towards our judgment-sensitive attitudes under different conditions (see Boyle, 2009).[17] We may collectively describe the experience as being rationally invested, as opposed to rationally indifferent, towards the thought in question. Thought insertion patients experience rational indifference towards theirs inserted thoughts in that they neither experience seamlessness nor the psychological tension caused by inconsistency or dissonance between their normative judgments and their inserted thoughts. In this section, I will

[17] This expectation is not what accounts for mine-ness when it comes to pains and other sensations and one might worry that my account does not provide a uniform analysis of mine-ness across all mental states. There are two points to make here: first, many self-knowledge theorists including Moran (2001), Carruthers (2011) and Boyle (2009) acknowledge a difference in the way we access the two types of mental states. My claim that we sense mine-ness differently in the two cases is consistent with, and an extension of, this line of thinking. Second, that the sense of mine-ness is differently constituted in the two cases coheres with the idea that our rational control over them differ significantly: while our sensory experiences are states that we must simply accommodate or tolerate irrespective of whether or not we can explain or understand them, we seem to have a far less passive role when it comes to desires, belies, and other attitudes.

provide four sources of support for my model: (a)patient reports, (b)findings from cognitive dissonance theory, (c)ordinary retrospective reflection and (d) reactions to counter-evidence.

5.1 Patient Reports

Two features of patient reports strongly support my proposed model. First, reports that describe the content of inserted thoughts suggest that they are judgment-sensitive:

> One evening the thought was given to me electrically that I should murder Lissi (Sousa and Swiney, 2013, 13).
> Thoughts crashed into my mind like a fusillade of rocks someone (or something) was hurtling at me- fierce, angry, jagged around the edges, and uncontrollable…I could not bear them, I did not know how to defend myself against them… *You are a piece of shit. You don't deserve to be around people…They can hurt you. They are powerful. You are weak.* (Saks, 2007, 83).

Note that such content can be sensibly subjected to normative scrutiny: why should Lissi be murdered? Why think that others are powerful and violent? Further, a subject's answers to such questions would, under ordinary circumstances, explain why she formed and maintained the corresponding beliefs.[18] This suggests that unlike sensations or 'brute' desires, inserted thoughts are judgment-sensitive. This is important because I have argued that our expectation of rational authority applies only to judgment-sensitive attitudes. Therefore, inserted thoughts are of the type that we would ordinarily expect to be rationally justified. This expectation would then be causally responsible for generating the impression that the thought is mine. Patients with thought insertion experience a loss of mineness precisely because they do not have such an expectation with respect to these judgment sensitive thoughts. Second, patients describe their relationship to their inserted thought in a way that expresses rational indifference:

> He said that he was getting 'queer ideas that are not of myself,' 'thoughts were given,' 'ideas that were not in my nature.' Subsequently he received mind suggestions, these came many times a day and dwelt on 'lewd low subjects'. He spoke of hypnotism and a greater power which 'came over me through this she-spirit operator.' The operators would throw lewd pictures into his mind. (Allison-Bolger, 2015, 237).
> [He] said […] 'it's like a thought as it comes in…a thought is very light really... it's a light feeling where you feel as though I'm actually thinking it…it's just a thought but it feels logical say… it feels pretty normal or fits with what I

[18] One might wonder here about spontaneous or unbidden thoughts. Such thoughts are often presented in propositional form, which means that they contain content that could be normatively examined in the way I've indicated here. They do not however appear to be judgment-sensitive, since their occurrence does not depend on the subject's rational considerations. I want to suggest, however, we do have an expectation of rational authority with respect to spontaneous thoughts in the sense that we expect to be able to *discard* such thoughts when they do not cohere with our normative considerations.

> suspect…I wonder if that's me…it felt like a piece of information…I don't think that was mine (Hoerl, 2001)

In the first report, the patient seems to disapprove of the lewdness of his inserted thoughts. Ordinarily, if we detected a thought we found distasteful we might react by dismissing the thought, or question our appraisal of it or perhaps revise our standards. If we cannot, we might feel ashamed or upset by its continued presence. However, this patient makes no effort to justify or render acceptable such thoughts. In addition, he does not express any psychological tension in being unable to justify his thoughts or integrate them within his normative framework—by, say, expressing guilt, shame or disappointment in himself for having objectionable thoughts, or confusion about why he would have such uncharacteristic thoughts. Rather, he matter-of-factly acknowledges their unfortunate and unavoidable presence. The second patient asserts that his inserted thoughts happen to resemble his own rational considerations, but despite the parallels, the patient does not experience seamlessness between his self-attribution and his inserted thought. His own normative assessment does not in and of itself give him automatic access to the content of his inserted thought. Despite the similarities between his own assessment and his inserted thought, he feels no automaticity or inevitability between his normative judgment and his inserted thought. Patients, then, seem to indicate an unmistakable sense of rational apathy toward the presence of their inserted thoughts.

Before moving on, I want to highlight a final feature of the condition indicated in these reports. Although patients feel rationally detached from their inserted thoughts, they appear to be intimately connected to them in a different way. Patients indicate that thoughts are "crashing into" their minds, "given to" them. If patients view such thoughts as entirely independent of their own rational considerations, what explains their experience that such thoughts are in fact occurring to *them*? I want to suggest that patients apprehend their inserted thoughts as judgment-insensitive sensory experiences. Depictions of inserted thoughts as *being received* or *coming in* or *crashing into* one's mind express the patient's sense that the thought, much like a toothache or a visual experience, is simply imposed on them and something that they must passively accommodate. Patients then feel rationally disconnected from their inserted thoughts, but nevertheless experience them directly.

5.2 Cognitive Dissonance Theory, Recalcitrant Attitudes and OCD

Findings in social psychology, and in particular, cognitive dissonance theory strongly support my claim that we ordinarily expect our conscious attitudes to be rationally justified. According to Leon Festinger, who is credited with introducing the theory, "the individual strives towards consistency" such that one's "opinions and attitudes… tend to exist in clusters that are internally consistent" (Festinger, 1957, 1). Festinger accepts that inconsistencies may nevertheless exist and persist, and when they do, such dissonance "leads to an activity oriented toward dissonance reduction just as hunger leads to activity oriented towards hunger reduction" (3). While the theory has undergone various revisions since its inception and has branched off in different ways from its initial formulation, there is general agreement among theorists that "people have a

desire to maintain consistency", and that awareness of inconsistency "arouses aversive feelings of dissonance" that in turn "trigger mental and behavioral reactions aimed at reducing these feelings" (Gawronski and Brannon, 2019, 101).[19]

A number of studies have focused on attitude change as the primary means by which to reduce aversive feelings associated with dissonance. But this can be difficult due to a variety of factors such as the importance of the attitude, (Stryzak et al., 2009; Devine et al., 2019) or because the cognitive element in question "is consonant with a large number of other elements and to the extent that changing it would replace these consonances by dissonances, the element will be resistant to change" (Festinger, 1957, 27). When this occurs, the subject might reduce discomfort by resorting to other strategies, such as recruiting additional cognitive elements that are consonant with a problematic attitude, or minimizing the importance of the conflicting elements. Without this recourse, subjects are left with some degree of psychological tension.

Research in cognitive dissonance supports my model in two ways. First, their findings directly support my claim that we generally expect our judgment-sensitive attitudes to be subjectively justified—in the sense of being internally coherent—and feel unsettled by discrepancy (particularly when the attitude or cluster of attitudes is meaningful to our lives). Second, such research also explains the experience of recalcitrant attitudes in a way that favors my proposal. According to my view, when we attribute recalcitrant attitudes to ourselves, we experience a pronounced discomfort, potentially leading to adjustments in our attitudes or evaluations. For example, the student who struggles with her belief in God in the face of counter-arguments might eventually come to devalue the significance of philosophical arguments against God's existence in order to accommodate her theistic belief. Such a process seems to be supported by findings in cognitive dissonance research. In one study, students who were strongly opposed to tuition increases were asked to write an essay in which they discussed the issue. Subjects who freely consented to write a *counter-attitudinal* essay reported greater discomfort than those who had consented to writing a pro-attitudinal essay (Elliot and Devine 1994). In addition, subjects who experienced discomfort retreated back to normal levels upon changing their attitude in the direction of the counter-attitudinal essay. In another study, subjects who did not change their attitude when asked to write a counter-attitudinal essay (this time concerning mandatory comprehensive final exams), instead trivialized the importance of one or more of the dissonant elements (Simon and Greenberg, 1995).

Such findings may also help explain the difference in the sense of ownership experienced by patients with obsessive-compulsive disorder (OCD) and thought-insertion patients. OCD patients report having intrusive and unwanted thoughts, some of which appear irrational to the patient, but which patients nevertheless acknowledge as their own (Levy 2003; Chamberlain et al., 2005; Meynen, 2012). They may be distressed by an inability to change or restructure their cognitive system in a way that removes the felt dissonance. Churchland (2002) writes "OCD patients often indicate that they wish to be rid of hand-washing or foot-step counting behavior, but cannot

[19] Some have modified Festinger's original theory by adding that the cause of cognitive dissonance and its reduction has to do with disruptions to one's self-image (Aronson, 1992), while others have argued that one must feel personally responsible for aversive consequences brought about before cognitive dissonance can mediate shifts in attitude (Cooper and Fazio, 1984).

stop" (208). Glannon (2005) similarly notes that such patients "feel that they must do certain things, or that they must think certain thoughts, though they claim that they do not want to have these feelings and thoughts and often desperately try to fight them" (73-4).[20] Unlike thought-insertion patients, sufferers of OCD are distressed by the disparity they perceive between their normative standards and their obsessive thoughts and attempt to alleviate their tension by changing their thought, that is, excising the problematic element. Since they cannot, they continue to be distressed by the presence of such thoughts, which is consistent with cognitive dissonance studies. This also explains the intact sense of ownership exhibited in OCD. The dissonance-induced distress constitutes their impression that such thoughts, while irrational, are their own.[21] In contrast, thought insertion patients are not particularly troubled by the fact that their inserted thoughts are at odds with their normative assessments. Recall the patient above who acknowledged having a thought at variance with his own standards, but did not feel ashamed or embarrassed, and did not have an immediate impulse or a sense of urgency to remove the unwanted thought.

5.3 Retrospective Reflection

That we ordinarily expect our conscious attitudes to be subjectively justified is also evidenced in our tendency to reconstruct our deliberative path to previously formed attitudes in a highly intelligible and justifiable manner. This is suggested by a number of studies which show that people tend to provide extremely rational portrayals of earlier intentions and beliefs. In Maier's 1931 experiment, subjects entered a room in which two cords hung from the ceiling of a laboratory that contained many other random objects such as poles, ring stands, clamps, pliers and extension cords. The subject's task was to tie the two ends of the cords together. The problem was that the cords were placed so far apart that the subject could not hold on to one cord while reaching the other. A short while after the experiment had begun, Maier, who was walking around the room would casually nudge one of the cords so that it was in motion, and within 45 seconds, the subject typically attached a weight to the end of one of the cords, swung it and ran to the other cord and waited for the first cord to swing close enough that it could be grasped. The interesting take away for our purposes is how agents explained their reasoning process post-hoc:

> It was the only thing left; I just realized the cord would swing if I fastened a weight to it. A psychology professor subject was more inventive. 'Having

[20] Some argue that obsessions begin with intrusive thoughts that are fairly common in the general population, but which OCD sufferers will appraise in distinctively faulty ways, by say, overestimating threats, assigning excessive importance to one's thought and an inflated sense of responsibility. Now, "once the intrusive thought is perceived in this highly exaggerated and threatening manner, the individual feels compelled to engage in escape, avoidance, compulsions or other forms of control responses" that will neutralize the anxiety associated with the intrusion or prevent the negative outcome (Clark and Guyitt, 2008). These reactions, while mildly successful, often increase their frequency and generate additional stress. Such theories also seem to support the idea that patients expect their obsessive thoughts to conform to their rational considerations and are distressed by the fact that they cannot.

[21] While one also experiences discomfort with recalcitrant attitudes, the distress suffered by OCD patients would not be the same, as they would differ in the extent and nature of the cognitive discrepancy, the subject's appraisal of the problematic thought and the frequency of their occurrence.

> exhausted everything else, the next thing was to swing it. I thought of the situation of swinging across a river. I had imagery of monkeys swinging from trees. This imagery appeared simultaneously with the solution. The idea appeared complete'. (Nisbett and Wilson, 1977, 241).

Researchers use this and similar studies to highlight the degree of introspective blindness or ignorance when it comes to our decision-making processes. But what is also noteworthy is the particular way in which subjects misinterpreted their process: subjects assumed that they arrived at their decisions by drawing from relevant past experiences and attending to contextual cues in a sensible and clever manner.

In another experiment, subjects whose minds were not entirely made up on a controversial and complex topic were questioned about their initial positions and then presented with a set of arguments either in favor of, or against, a position on the topic. Then the experimenter measured people's attitudes again, which now turned out to be closer to the persuasive message that they were exposed to. Participants were then asked to describe their *former* positions. This is where it becomes interesting: people "[retrieved] their current ones instead—an instance of substitution—and many [could not] believe that they ever felt differently" (Kahneman, 2011, 202). Notice what occurred: in reflecting on their former belief, subjects assumed that they had always accepted what seemed to them the most reasonable option.

A final example comes from research on split-brain patients, where communication between the two hemispheres is prevented because of the severed corpus callosum. A picture of a snow-filled meadow is presented to the non-verbal right hemisphere, while a picture of a chicken's claw is presented to the verbal left hemisphere. The patient is then asked to select from a range of pictures which one goes with the stimuli he was previously presented. The patient typically selects two pictures: the right hemisphere leads the left hand to select a shovel, while the right hand selects a chicken to go with the claw originally presented to the left hemisphere, which are both fitting to the two pictures the patients were initially shown. What pertains to our purposes happened next: when patients were asked to explain their choices, they offered responses such as "The chicken claw goes with the chicken and you need a shovel to clean out the chicken shed" (Gilovich, 1991, 22). The real reason the patient pointed to the shovel (the picture of snow) was not given because that information was not communicated to the left hemisphere in which language ability is localized. Nevertheless, the patient gave a "sensible response" and "[invented] a story to account for it" that "easily [made] sense of even the most bizarre patterns of information" (Gilovich, 23).

When asked to reflect on their former beliefs or intentions, subjects in all these cases tended to favor a reasonable and intelligible cognitive path. These examples illustrate a tendency to assume that our attitudes are, by and large, rational.

5.4 Addressing Counter-Evidence

We exhibit a strong propensity to actively seek ways to neutralize potential counter-evidence against out beliefs. Upon confronting data that would, if accepted, rationally demand that we abandon our existing attitude, we do not typically blatantly ignore such evidence or maintain our attitude and accept the evidence. Instead, we take efforts to

convince ourselves that the evidence does not actually undermine our convictions. Westen et al., (2006) conducted a study in which subjects who rated themselves strong Democrats or Republicans were shown clips containing claims made by John Kerry or George Bush. They were then shown clips (some of them fictional) in which the two political figures contradicted themselves. Next, subjects were asked to consider whether the targets' statements and actions were inconsistent with each other. Then, subjects were shown an exculpatory statement that logically explained away the inconsistency, and asked whether the political target's statements and actions were not quite as inconsistent as they first appeared. Perhaps unsurprisingly, the experimenters found that "partisans were substantially more likely to accept the exculpatory statements for their own candidate versus the opposing candidate" (Westen et al., 2006, 1950). Participants, in other words, "recruited beliefs that allowed them to resolve what others would see as an obvious contradiction" (Lynch, 2012, 13).

Thomas Kelly (2008) has also discussed this phenomenon with respect to attitude polarization, noting that when parties who disagree are given evidence against their view, rather than revising their original judgments, they will each "harden his or her opinion and the gulf.. widens" (612). While this may appear irrational and imply a willingness to retain conflicting cognitive elements (one's belief and counter-evidence against the belief), Kelly explains that there is in fact more going on. He argues that a strong believer in *p* who encounters what appears to be a sound argument for *not-p* "is disposed to view that argument with a greater measure of suspicion and subject it to closer scrutiny. And the more one subjects the argument to close scrutiny, the more likely one is to find a flaw in that argument if in fact there is some flaw to be found" (Kelly, 2008, 618). Notice here that one is not merely ignoring the evidence or adamantly believing something for which one has no rational support. Rather, people react in a way that allows them to sustain justification for their view.[22] This is supported by Lord et al.'s (1979) study in which subjects were given data that conflicted with their views on the death penalty. Gilovich (1991) reports that participants

> carefully scrutinized the studies that produced these unwanted and unexpected findings, and came up with criticisms that were largely appropriate. Rather than ignoring outright the evidence at variance with their expectations, the participants cognitively transformed it into evidence that was considered relatively uninformative and could be assigned little weight (Gilovich, 1991, 54).

What is notable in these cases is that subjects do not typically insist on a belief that they take to be straightforwardly at odds with the available evidence, but rather reinterpret the data in a way that renders their existing belief justifiable to them. Again, this supports my view that we generally expect our conscious beliefs and other attitudes to be rationally justifiable to us. More generally, I have attempted to defend two claims in this section: first, patient reports suggest that patients lack an expectation of rational authority towards their inserted thought which produces the sense that the thought is not their own, an experience constituted by rational apathy towards their thought. Second, I've offered empirical evidence from cognitive and social psychology that

[22] Kelly also maintains that a different mechanism by which we respond to counter-evidence is to search for alternative explanations for evidence that supports a hypothesis that we antecedently reject.

suggests that we ordinarily have an expectation of rational authority towards our attitudes.

6 Objections

6.1 Attributing Irrational Attitudes without Psychological Discomfort

I argued above that ordinarily one experiences the expectation of rational authority towards their judgment-sensitive states in two ways. When the expectation is met, there is a seamless transition between one's normative assessments and the states one attributes to oneself. When the expectation is not met, the subject experiences a form of psychological tension that prompts adjusting of one's cognitive network to alleviate the discomfort. However, one might worry that my account over-intellectualizes the relationship one generally has to one's thoughts. While we expect *some* of our beliefs—perhaps those that require careful reflection, such as the best time to put one's house on the market, or whether or not to promote an employee — to depend on our reasoning, we do not apply the same standard to many of our other, more casual beliefs, and instead, accept with resignation that we are not fully rational creatures with fully justifiable attitudes.[23] We might imagine a subject who finds that she has a desire for a romantic relationship with an individual who is undependable and untrustworthy. Despite the foreseeable perils, she resigns herself to accepting that this is her desire, and that she must simply follow her heart, the objects of which her mind cannot comprehend. Or take the case of a religious devotee who believes that her theistic beliefs are not rationally justified, but unhesitatingly remains committed to her belief in God because she believes that human cognitive faculties are ultimately incapable of comprehending the divine. These cases challenge my own view in light of the fact that subjects view these attitudes as their own, take them to be rationally *unjustified*, but experience no discomfort. They feel no pressure to adjust their cognitive system in a way that renders their attitude rationally justified from their perspective. This raises some doubt about the existence of an underlying unmet expectation of rational authority.

A different reason to doubt the existence of an expectation of rational authority surfaces in the self-knowledge literature and, in particular, as a criticism against Moran's conception of ordinary self-knowledge. Lawlor (2009) argues that in some cases, one comes to know one has a desire by attending to, and drawing an inference based on, one's *internal* promptings. For example, a mother who attempts to determine whether or not she wants another child "catches herself imagining, remembering, and feeling a range of things" in her daily life. These experiences "may be enough to prompt [her] to make a self attribution… Saying 'I want another child', she may feel a sense of ease or settledness" (57). She might instead give her feelings and memories more thought and "shift from the passive experience of fleeting imaginings to a more active prompting of her imagination" (58) asking herself why she keeps thinking about it, which might gradually lead to a self-attribution of one sort or another. Notice here

[23] Thanks to participants at the Southern Society of Philosophy and Psychology 2019 conference for raising this point.

that there is no assumption on the mother's part that her desire will be *rationally* justifiable. That is, she has not attempted to, nor is their any indication that she will be able to, determine why having another child might be generally worthwhile or desirable.

Do such considerations speak against an expectation of rational authority? The principal suggestion above is that it not atypical to lack such an expectation towards one's thoughts: we unproblematically accept that parts of our mental economy will not be justifiable and we apprehend our own attitudes without a presumption of rationality.

Having an expectation of rational authority toward my attitude *a* does not consist merely in the expectation that *a* will be justified by my understanding of facts in the outside world, facts that are external to me. Rather, the expectation is that *a* will be justified by the types of reasons that I take to be relevant to *a* and the way in which I rank the relative importance of each reason. In the case of attitudes having to do with one's personal or social life, such as having more children or starting a romantic relationship, one might well consider facts about how the attitude affects one's existing mental life as legitimate reasons for or against the attitude. We can imagine the religious devotee, for example, minimizing the weight of impersonal, philosophical arguments and inflating the importance of the positive cognitive and emotional impact of her belief. She might call to mind the sense of purpose and reassurance afforded by her belief in god, the sense of community she receives from interacting with other believers, and so on. The mother who imagines life with a newborn might observe that the prospect of having and caring for another child does not strike her as overwhelming, burdensome or daunting, but rather, feels exciting and profoundly fulfilling and might rank such considerations as the most important justification for the desire.

What we see in these cases is that subjects do not experience psychological tension or discomfort because their self-ascribed attitudes *do* meet their expectation of rational authority. Their attitudes are subjectively justified in that they are supported by the sorts of reasons that the subject considers most relevant to the attitude. Subjects might casually or loosely describe such attitudes as irrational, but this is because they ignore the fact that they justify these attitudes in light of how they affect their own mental lives.

6.2 Automatic, Non-reflective States

A different reason to doubt the existence of such an expectation is that many of our thoughts, even those that appear judgment-sensitive, are automatic and produced by quick, uncritical and non-reflective processes. Kahneman (2011), in *Thinking, Fast and Slow,* distinguishes two systems in the mind: System 1 generates judgments, feelings and inclinations rapidly and with little effort, while system 2 operates carefully, slowly and critically. System 1 effortlessly produces impressions, feelings and often complex judgments that are the main sources of the deliberate beliefs and choices of system 2. If our primary mode of thinking consists in the operation of system 1, Cassam maintains that "much of the time, our reasoning isn't guided by an appreciation, use and assessment of reasons and reasoning as such." (2014, 17).[24] Now, if it is true that we

[24] Cassam clarifies that such short-cuts in our thinking are not necessarily ill-founded or irrational, as they often allow us to efficiently navigate the various demands made on our limited cognitive resources.

form most of our beliefs without careful thought and consideration, the idea that we expect our beliefs and other attitudes to be rationally supportable starts to look less likely.

While system 1 operates without deliberate or active effort, we nevertheless expect its deliverances to conform to our rational assessments.[25] This is evidenced by the fact that when system 1 "runs into difficulty" or a "question arises for which system 1 does not offer an answer", the careful and detailed processing of system 2 is mobilized (Kahneman, 24). This suggests that the products of system 1 are subjected to rational monitoring and may be discarded, questioned or revised when they conflict with one's own normative considerations. Individuals accept system 1's outputs to the extent that they are consistent with their expectations.[26]

There are a number of ways in which system 1 generates the impression that one's judgment is consistent and justified, despite not involving active deliberation. For example, in contexts where there is some ambiguity—say the statement 'Mary went to the bank', the subject does not explicitly or deliberately sort through alternative interpretations (a financial institution, a water's edge). Rather, system 1 automatically makes an interpretive choice based on recent experiences and contextual cues. Because "system 1 does not keep track of alternatives that it rejects, or even of the fact that there were alternatives", the subject is led to view her immediate understanding as plainly correct, experiencing no "conscious doubt" or "subjective discomfort" (Kahneman, 80). In addition, system 1 is responsible for what psychologists call the halo effect. Here, subjects go from knowing one or two positive attributes of an individual to immediately believing that the individual has several other unrelated positive attributes. This illustrates "one of the ways the representation of the world that system 1 generates is simpler and more coherent than the real thing" (82). We maintain these initial and automatic impressions so long as our experiences and judgments continue to support them. When we detect a feature that fails to cohere with our expectations and standards, we summon the more careful and explicit processing of system 2. In essence, system 2 is "mobilized to increased effort when it detects an error" in the fluid and automatic outputs of system 1 (25). Thus, the absence of active and deliberate mental activity in system 1 does not entail the absence of an expectation of rational authority.

6.3 Expectation of Rational Authority Is Maintained

I've argued that thought-insertion patients lack an expectation of rational authority over their inserted thoughts, which is experienced as rational indifference toward such

[25] The products of system 1 are diverse, and include innate inclinations and learned associations that "become fast and automatic through prolonged practice" and "broad knowledge of the language and the culture" (Kahneman, 22). Some of these are judgment-sensitive insofar as it makes sense to subject them to normative questioning.

[26] Kahneman divides reliance on System 1 into good and bad cases. System 1 can generate unwarranted confidence by inventing a causal story based on scarce data or forge associations when there are none to be found. However, when System 1 is trained within a predictable environment and the subject has adequate opportunity to register the patterns salient in that environment, then the automatic deliverances of System 1 can be trusted. Irrespective of the actual accuracy and reliability of System 1, subjects feel as though the judgments propagated here are perfectly justified in light of the degree of coherence generated by the judgments, and the ease with which they are accepted.

thoughts. One might worry, however, that this model does not account for the fact that patients are distressed by the presence of such thoughts. Doesn't this suggest an unmet expectation?[27,28]

While some patients do appear troubled by the presence of such thoughts, their distress results not from an unmet expectation, but rather a dual impression of the thought.

Patients view their inserted thought as, on the one hand, judgment-sensitive, but on the other, as entirely impervious to their own rational considerations. That is, they realize that their thought would ordinarily warrant an expectation of rational authority, but they have no such expectation towards the thought. This dual realization is captured in Saks's (2007) autobiography:

> Once there'd been a time in my life when thoughts were something to be welcomed, and pored over, like pages in a favorite book… But now thoughts crashed into my mind like a fusillade of rocks someone (or something) was hurtling to me—fierce, angry, jagged around the edges, and uncontrollable. I could not bear them, I did not know how to defend myself against them. (83)

There is a recognition, on the one hand, that thoughts are the sort of thing that one enjoys and can exert control over, but she is struck by the fact that she does not enjoy the same authority over her inserted thoughts. Rather, they intrude upon her mind and remain there without being solicited or supported.

There are two reasons to think that this is the source of patient distress. First, patients with thought insertion do not exhibit confusion about the meanings of ordinary words, and do not employ their own private, idiosyncratic definitions, as is the case with other symptoms of schizophrenia (Sims, 2003). Liddle's (1987) classification of schizophrenic symptoms (a classification that has since gained wide acceptance) (Frith and Johnstone, 2003) presents a tri-fold division: (a)reality distortion, which includes hallucinations and delusions, (b)disorganization, which includes incoherent speech and inappropriate emotional responses, and (c)psychomotor poverty, which consists of poverty of speech, poverty of action and blunted emotional responses. While these do not correspond to subgroups of patients, they do present clusters of symptoms, such that "a patient who presents with hallucinations and delusions does not necessarily tell us anything about whether or not the patient will show poverty of speech or disorganized behavior" (Frith, 62). In Liddle's (1987) study, patients were divided into three groups using this syndrome approach. Those who exhibited reality distortion showed little evidence of cognitive dysfunction, unlike those who exhibited disorganization and psychomotor poverty, who performed badly in tests of long-term memory, conceptual thinking, immediate recall and word learning. Thought insertion is more generally regarded as a positive symptom, so called because its abnormality lies in

[27] Thanks to both Patty McShane and audience members at the Southern Society for Philosophy and Psychology 2016 meeting, for raising this worry.

[28] It is important to note that unlike psychiatric patients who suffer from affective disorders (such as depression or anxiety), delusional patients do not typically sense that anything is wrong. Patients diagnosed with schizophrenia do not view their hallucinations and delusions as pathological symptoms, but rather, accurate representations of their surroundings.

its presence, and not by the "absence of normal functions, as is the case with reduced speech output" (Fletcher and Frith 2009, 48). Both psychologists and philosophers have regarded thought-insertion patients as providing meaningful accounts of their experiences: "When one reads the reports, patients seem to be describing their experiences in a meaningful way" (Fernandez 2013, 14); "These are compelling reports of experience which many people are giving; at the very least we should want to understand why it is so natural, so compelling to describe experience in this way" (Campbell, 1999, 610). These classifications and interpretations of the condition suggest that patients will not have trouble recognizing a judgment-sensitive attitude: they will recognize it as the type of attitude that ought to be constrained by one's own rational deliberations. Yet, they are struck by the sort of rational indifference they experience towards the thought.

Second, even when patients report that the content of the inserted thought aligns with their own assessments, this fact does not alleviate their discomfort. This suggests that in thought insertion, what troubles the patient is not that their thought does not match their normative considerations, but rather that they do not *expect* such thoughts to depend on, or be shaped by, such considerations.

6.4 Non-Explanatory Account

A final concern is that my account does not explain the experience of thought insertion, but seems to be read more plausibly as a consequence of the experience.[29]

That is, it seems natural to think that because a thought does not feel like one's own, one will not expect to have rational authority over that thought. While the direction of explanation outlined in the objection might appear natural, the intuition is based on an erroneous assumption, namely that we can make sense of the experience of dis-ownership without further explication. But what needs explaining is this very thing: how one can lack a sense of mineness when it comes to introspectively accessed thoughts? What I've suggested is that the experience of dis-ownership amounts to rational indifference which is itself produced by an absence of the expectation of rational authority. This allows us to make sense of the patient's claim that the thought does not feel like her own.

7 Conclusion

I have argued here that current SK models fail to provide an adequate account of the experience of thought insertion, and in particular, the experience of dis-ownership. Attempting to characterize the experience as an inability to endorse one's inserted thought confronts the problem of recalcitrant attitudes, thoughts recognized as one's own, despite an inability to justify them. I have proposed and attempted to defend the view that thought-insertion patients do not expect their inserted thoughts to be rationally justified, which then produces an experience of rational indifference towards their inserted thought.

[29] Thanks to Andrew Eshleman and Bradley Griggs for raising this question.

References

Aronson, Elliot. 1992. The return of the repressed: Dissonance theory makes a comeback. *Psychological Inquiry* 3 (4): 303–311.

Bolger, Allison. 2015. Locating thought insertion on the map of ordinary thinking. *Philosophy, Psychiatry & Psychology* 22 (3): 235–238.

Bortolotti, Lisa, and Matthew Broome. 2009. A role for ownership and authorship in the analysis of thought insertion. *Phenomenology and Cognitive Sciences* 8: 205–224.

Boyle, Matthew. 2009. Two kinds of self-knowledge. *Philosophy and Phenomenological Research* 78 (1): 133–164.

Campbell, John. 1999. Schizophrenia, the space of reasons and thinking as a motor process. *The Monist* 82 (4): 609–625.

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

Cassam, Quassim. 2014. *Self-knowledge for humans*. Oxford: Oxford University Press.

Chamberlain, S., A. Blackwell, N. Fineberg, T. Robbins, and B. Sahakian. 2005. The neuropsychology of obsessive compulsive disorder: The importance of failures in cognitive and behavioral inhibition as candidate endophenotypic markers. *Neuroscience and Biobehavioral Reviews* 29: 399–419.

Churchland, P. 2002. *Brain-wise: Studies in Neurophilosophy*. Cambridge: The MIT Press.

Clark, David, and Brendan Guyitt. 2008. Pure obsessions: Conceptual misnomer or clinical anomaly? In *Obsessive compulsive disorder: Subtypes and Spectrum conditions*, ed. Jonathan Abramowitz, Dean McKay, and Steven Taylor, 53–75. New York: Elsevier.

Cooper, and Fazio. 1984. A new look at dissonance theory. In *Advances in experimental social psychology*, ed. L. Berkowitz, 229–264. Orlando: Academic Press.

Devine, P., J. Tauer, K. Barron, A. Elliot, K. Vance, and E. Harmon-Jones. 2019. Moving beyond attitude change in the study of dissonance-related processes: An update on the role of discomfort. In *Cognitive dissonance: Reexamining a pivotal theory in psychology*, ed. E. Harmon-Jones, 247–269. Washington, DC: American Psychological Association.

Elliot, A., and P. Devine. 1994. On the motivational nature of cognitive dissonance: Dissonance as psychological discomfort. *Journal of Personality and Social Psychology* 67 (3): 382–394.

Fernandez, Jordi. 2013. *Transparent minds*. Oxford: Oxford University Press.

Fernandez, Jordi. 2010. Thought insertion and self-knowledge. *Mind and Language* 25 (1): 66–88.

Festinger, L. 1957. *A theory of cognitive dissonance*. Evanston: Row, Peterson and Company.

Fletcher, P., and C. Frith. 2009. Perceiving is believing: A Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews Neuroscience* 10: 48–58.

Fricker, Miranda. 2007. *Epistemic injustice: Power and the ethics of knowing*. Oxford: Oxford University Press.

Frith, Christopher, and Eve Johnstone. 2003. *Schizophrenia: A very short introduction*. Oxford: Oxford University Press.

Frith, C. 1992. *The cognitive neuropsychology of schizophrenia*. Hillsdale: Erlbaum.

Gawronski B. and Brannon, S. 2019. What is cognitive consistency, and why does it matter? Cognitive dissonance: Reexamining a pivotal theory in psychology, ed. E. Harmon Jones, 91-116. Washington, DC: American Psychological Association.

Gilovich, T. 1991. *How we know what isn't so: The fallibility of human reason in everyday life*. New York: The Free Press.

Glannon, W. 2005. Neurobiology, neuroimaging, and free will. *Midwest Studies in Philosophy* 29: 68–82.

Graham, George, and Lynn Stephens. 1994. Mind and mine. In *Philosophical psychopathology*, ed. George Graham and Lynn Stephens, 91–109. Cambridge: MIT Press.

Hoerl, Christopher. 2001. On thought insertion. *Philosophy, Psychiatry and Psychology* 8 (2/3): 189–200.

Kahneman, D. 2011. *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.

Kelly, T. 2008. Disagreement, dogmatism, and belief polarization. *The Journal of Philosophy* 105 (10): 611–633.

Lawlor, K. 2009. Knowing what one wants. *Philosophy and Phenomenological Research* 79 (1): 47–75.

Levy, Daniel. 2003. Neural holism and free will. *Philosophical Psychology* 16 (2): 205–228.

Liddle, P. 1987. Schizophrenic syndromes, cognitive performance and neurological dysfunction. *Psychological Medicine* 17: 49–57.

Lord, C.G., L. Ross, and M. Lepper. 1979. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology* 37: 2098–2109.

Lynch, M. 2012. *In praise of reason*. Cambridge: The MIT Press.

Maier, N. 1931. Reasoning in humans II: The solution of a problem and its appearance in consciousness. *Journal of Comparative Psychology* 12: 181–194.

Mellor, C. 1970. First rank symptoms of schizophrenia. *British Journal of Psychiatry* 117: 15–23.

Meynen, G. 2012. Obsessive-compulsive disorder, free will and control. *Philosophy. Psychiatry and Psychology* 19 (4): 323–332.

Moran, Richard. 2012. Self-knowledge, 'transparency' and the forms of activity. In *Introspection and consciousness*, ed. Declan Smities and Daniel Stoljar, 211–236. Oxford: Oxford University Press.

Moran, Richard. 2001. *Authority and estrangement*. Princeton: Princeton University Press.

Nisbett, R., and T. Wilson. 1977. Telling more than we can know: Verbal reports on mental processes. *Psychological Review* 87 (3): 231–259.

Parrot, Matthew. 2017. Subjective misidentification and thought insertion. *Mind & Language* 32 (1): 39–64.

Pickard, Hanna. 2010. Schizophrenia and the epistemology of self-knowledge. *European Journal of Analytic Philosophy : Special Issue in Classification and Explanation in Psychiatry* 6: 55–74.

Saks, Elyn. 2007. *The Center cannot hold: My journey through madness*. Boston: Hachette Books.

Scanlon, T.M. 1998. *What we owe to each other*. Cambridge: Belknap Press of Harvard University Press.

Simon, L., and J. Greenberg. 1995. Trivialization: The forgotten mode of dissonance reduction. *Journal of Personality and Social Psychology* 68 (2): 247–260.

Sims, A. 2003. *Symptoms in the mind: An introduction to descriptive psychopathology*. Edinburgh: Elsevier Science Limited.

Sollberger, Michael. 2014. Making sense of an endorsement model of thought insertion. *Mind and Language* 29 (5): 590–612.

Sousa, Paulo, and Lauren Swiney. 2013. Thought insertion: Abnormal sense of thought agency or thought endorsement? *Phenomenology and the Cognitive Sciences* 12 (4): 637–654.

Westen, D., P. Blagov, K. Harenski, C. Kilts, and S. Hamann. 2006. Neural bases of motivated reasoning: An fMRI study of emotional constraints on partisan political judgment in the 2004 US presidential election. *Journal of Cognitive Neuroscience* 18 (11): 1947–1958.