palgrave communications
HUMANITIES | SOCIAL SCIENCES | BUSINESS

ARTICLE

Received 24 Aug 2016 | Accepted 19 Dec 2016 | Published 24 Jan 2017

DOI: 10.1057/palcomms.2016.108 OPEN

Kant and thought insertion

Sacha Golob[1]

ABSTRACT This article examines the phenomenon of thought insertion, one of the most extreme disruptions to the standard mechanisms for self-knowledge, in the context of Kant's philosophy of mind. This juxtaposition is of interest for two reasons, aside from Kant's foundational significance for any modern work on the self. First, thought insertion presents a challenge to Kant's approach. For example, the first Critique famously held that "The 'I think' must be able to accompany all my representations" (Kant, KrV, B132). Yet thought insertion raises the problem of representations that are "mine" by many natural criteria, and yet which I am unwilling to self-ascribe. Ultimately, my argument will be that thought insertion simultaneously problematises, and yet to some degree also vindicates, the complex distinctions between activity and passivity that underlie Kant's system. Second, I argue that Kant's position contains resources that allow us to model thought insertion, and its broader implications for self-knowledge, in an interesting and distinctive manner. Kant himself held an extreme view of philosophy's competence in the study of mental disorder: in the Anthropology, he suggests that courts must refer such cases to philosophers, rather than medics (Kant, Anth, p. 214). My aim is much more modest: to suggest that a Kantian treatment of thought insertion deserves consideration by both philosophers and clinicians. This article is published as part of a collection on self-knowledge in and outside of illness.

[1] Philosophy, King's College London, London, UK Correspondence: (e-mail: sacha.golob@kcl.ac.uk)

Introduction

This article examines the phenomenon of thought insertion, one of the most extreme disruptions to the standard mechanisms for self-knowledge, in the context of Kant's philosophy of mind.[1] This juxtaposition is of interest for two reasons, aside from Kant's foundational significance for any modern work on the self. First, thought insertion presents a challenge to Kant's approach. For example, the first *Critique* famously held that "The 'I think' must be able to accompany all my representations" (Kant, KrV, B132).[2] Yet thought insertion raises the problem of representations which are "mine" by many natural criteria, and yet which I am unwilling to self-ascribe. Ultimately, my argument will be that thought insertion simultaneously problematizes, and yet to some degree also vindicates, the complex distinctions between activity and passivity which underlie Kant's system. Second, I argue that Kant's position contains resources that allow us to model thought insertion, and its broader implications for self-knowledge, in an interesting and distinctive manner. Kant himself held an extreme view of philosophy's competence in the study of mental disorder: in the *Anthropology*, he suggests that courts must refer such cases to philosophers, rather than medics (Kant, Anth, p. 214). My aim is much more modest: to suggest that a Kantian treatment of thought insertion deserves consideration by both philosophers and clinicians.

Before proceeding, a few remarks on scope. This article is not intended primarily as exercise in textual Kant scholarship. This is partly because pinning down the views of the historical Kant would require close exegetical discussion of many connected areas: for example, his complex stance on rationalist models of the self, or on the general relationship between understanding and intuition. My aim is rather to show how a broadly Kantian approach can accessibly and helpfully contribute to the existing discussion of this complex medical phenomenon. In line with this, I will try to keep Kantian technical terms to a minimum. I am also not going to attempt to do justice to the full range of the existing philosophy of mind literature: for example, I will not discuss recent work on thought insertion and inner speech.[3] The underlying assumptions in play there—with respect to the nature of such speech, its frequency and its links to ideas such as Evansian transparency—are too different from Kant's own for a fruitful discussion to be possible in the current context. Finally, I will treat the "thoughts", which are inserted as beliefs (I say much more on what this amounts to below). There are interesting case of what are prima facie "emotion insertions" in patient reports, but, given the emphasis on reason-responsivity which characterizes the Kantian approach, these present a very different set of challenges. I briefly indicate why in the conclusion, but a full treatment is beyond the present article.[4]

Definitions and the basic Kantian apparatus

This section will introduce and define the basic concepts I will use. Kant was an enormously systematic thinker, and the ideas I isolate are, within his own work, embedded in a perhaps uniquely complex larger system. I cannot do justice to this here, and I will, in line with the aims set out above, simply bracket it.[5]

Thought-insertion is a delusion whereby "the subject believes that thoughts that are not his own have been inserted into his mind" (Andreasen, 1984: 3). What is striking is that while subjects retain introspective access to the relevant thoughts, they are insistent that they are not theirs: the experience is of thoughts "which are not [their] own intruding" into their consciousness (Wing *et al.*, 1974: 160). Thought insertion so defined should be distinguished from two potentially accompanying phenomena. First, patients typically offer an explanation as to how this has come about; for example, because of electrical transmission (Jaspers, 1963: 580). Such explanations, whilst manifestly problematic, do not carry the hallmark feature of alienation: the patient regards the beliefs that articulate the explanation as their own, and so I will not discuss this aspect of the situation here. Second, the core phenomenon of thought insertion may be supplemented by a further belief as to whom the relevant thoughts actually belong. For example, this is a widely cited case from Mellor:

> I look out of the window and I think the garden looks nice and the grass looks cool, but the thoughts of Eamonn Andrews come into my mind. There are no other thoughts there, only his…He treats my mind like a screen and flashes his thoughts onto it like you flash a picture. (Mellor, 1970: 17)

One might thus contrast a "negative element" (that is, the fact that the patients don't recognize certain thoughts as belonging to them) with this further "positive element" (that is, the fact that they ascribe those thoughts to others) (Hoerl, 2001: 190). But this further element is not necessarily present (for example, Jaspers, 1963: 580), and I will not focus on it here—although I suggest in §4 that my approach can make sense of it.

The basic characterization of thought insertion that I have given is a simple one, and it is natural to make it more precise by introducing notions such as ownership. But these terms, and their inter-relations, are highly contested: there are approaches that present the subject as retaining ownership whilst lacking some further factor (Gerrans, 2001), and approaches which would deny even that (Bortolotti and Broome, 2009). So rather that importing such terminology at this point, I want first to introduce the Kantian apparatus I will use to frame my own account.

The place to begin is his conception of self-knowledge. There are, of course, many senses in which I might have knowledge about myself. But the one on which philosophers have focussed and the one relevant here is knowledge of our own mental states, a knowledge that is often thought to have distinctive characteristics of immediacy and authority. Kant draws a foundational distinction between two forms of such knowledge, a distinction which he articulates in terms of an active/passive contrast. On the one hand, there is what he calls "inner sense"—this is "a mere faculty of perception", an awareness of "what happens *to us*", such as feeling pain from an injury (Kant, Anth, 153,161). On the other hand, there is pure or transcendental apperception: whereas inner sense is consciousness of what affects us, apperception is consciousness of "what the human being *does* … this belongs to the faculty of thinking (Kant, Anth, p. 161).[6] Kant's shorthand for this active or "spontaneous" capacity for thought is the "I think" (Kant, KrV, A107, B132).

The idea that thought insertion can be usefully analysed in terms of some kind of active/passive framework is familiar from many standard accounts: the inserted thoughts would be something "done to" the patient, rather than things "done by" her (Fulford, 1989: 220). But what makes Kant distinctive is the *way* he understands the relevant notion of activity, and it is this I want now to discuss. I will approach the issue via the characteristically Kantian theme of self-determination.

There are many senses in which we might be said to determine ourselves. Not all of these are on the same footing. Suppose John takes steroids to speed up his muscle growth. The reflexive structure of this act is purely incidental—he could equally have used the same drugs to speed up someone else's development. Moran puts the underlying point nicely:

> In various cases a person may produce in himself various desires, beliefs, or emotional responses, either by training, mental discipline, drugs, the cooperation of friends, or simply

> by hurling himself into a situation that will force a certain response from him….[T]he resulting attitude is still one I am essentially passive with respect to. It is inflicted on me, even if I am the one inflicting it. (Moran, 2001: 117)

Suppose alternately that I am faced with some factual question; I consider the evidence, deliberate, and thereby acquire the belief that *P*.[7] This ability, an ability to exercise a rational authority over one's own beliefs, seems very different from the steroid case:

> [T]here is surely an intuitive contrast between my power to govern whether I have a stomach ache and my power to govern whether I believe *P*: whereas in the former case my control over the relevant condition is at best indirect, in the latter, one wants to say, my control may be direct. It is this intuition—that settling on an answer to a question can itself be an exercise of some sort of capacity for self-determination—that is expressed in the traditional idea that rational creatures have a capacity for free "judgment", a capacity to "make up their minds". (Boyle, 2011: 17)

The basic idea is thus that rational agents are able to exercise a particular form of self-regulation: I determine what my belief is by establishing what is *to be believed*, given the facts before me. As Moran has argued, this distinctive form of self-determination can be used to underwrite a distinctively active story about self-knowledge, one on which I have special authority to know what I believe, precisely because it is I who determine what it is I believe —not in the voluntaristic sense that I can pick my views at random, but in the sense that it is my mind to make up. As he puts it:

> If it is possible for a person to answer a deliberative question about his belief at all, this involves assuming an authority over, and a responsibility for, what his belief actually is. Thus a person able to exercise this capacity is in a position to declare what his belief is by reflection on the reasons in favor of that belief, rather than by examination of the psychological evidence. In this way … avowal can be seen as an expression of genuine self-knowledge (Moran, 2004: 425).

This has the attractive feature of meshing neatly with what is often called the transparency principle, the fact that when asked whether I believe that *P*, I typically attend to the external world, rather than "looking inward": in Evan's (1982: 225) famous formulation, when asked whether I believe that there will be a third world war, I consider the geopolitical situation rather than engaging in some kind of introspection. On Moran's approach this makes immediate sense: I can answer questions about at least reason-responsive mental states, such as beliefs, by looking at the world because my verdicts on the latter constitute the former.

Moran's influential work has faced multiple challenges in the literature—for example, I might sincerely judge that *P* and yet still not form the belief that *P* when belief is understood to include some rich set of dispositions (consider the phenomenon of implicit bias).[8] But in the present context, my interest is not in a general assessment of Moran (for my views on this see Golob, 2016), but rather in two specific points. On the one hand, Moran's approach serves as a good example of the type of Kantian approach I will sketch: this is because it powerfully illustrates the idea that rational agents are able to exercise a specific kind of self-regulation. On the other hand, however, as has been widely discussed and as is visible even in my quick summary, Moran naturally emphasizes cases in which I generate beliefs through explicit deliberation. This has problematic consequences. For example, it means that the theory is less attractive when the relevant question is not "do you believe that *P*?" but rather "do you already believe that *P*?"—since the latter is naturally read not as asking me to now form a view on the facts, but rather to report some antecedently existing state of affairs (Shah and Velleman, 2005). This type of concern meshes closely with broader post-Kantian worries about Kant's talk of apperception or self-consciousness: how is this to be understood given that the vast majority of our behaviour does not involve thematic, explicit reflection?[9] So what I want to do is to press the idea of rational self-governance further, but shift the focus away from conscious deliberation and towards the distinctive forms of behaviour, thematised or unthematised, which rationality makes possible.

For Kant one of the key features of rational agents such as humans, in contrast to non-rational animals (henceforth "animals"), is an ability to "recognise" the "marks" or properties of the things they encounter (Kant, SvF p. 59). On his view, while animals will obviously respond differently depending on the properties of the world around them (a dogs recoils from a heated piece of metal, but not a cool one), only rational agents are able to represent the fact that certain properties imply certain others. This recognitional capacity is what defines concept use for Kant, and he therefore cashes concepts in terms of "rules": for example, to possess the concept of <body> is to represent the fact that its application in turn "necessitates the representation of extension" (Kant, KrV A106, A126).[10] To adapt Kant's own example, an ox may respond very differently to a stall made of paper and one made of steel; but only a rational being is able to represent the fact that its being paper implies, given the relevant background conditions, other properties (flammability, for example). It is in this sense that the *Logic* treats marks as both "in the thing" and as a "partial representation … considered as the ground of cognition" (Log.:58): the recognition of such properties regulates our representations.

Why does this matter? The answer is that for Kant this capacity transforms the way in which rational beings engage with themselves, with each other and with the world. Let me give two examples. At an individual level, such recognition brings my representations within the domain of a specific kind of assessment. On Kant's picture, the ox simply sees one thing after another—it lack spontaneity in Kantian terms (I discuss association shortly). Rational beings, in contrast, are continually taking on commitments: to see the door as wooden means that I must be willing to affirm various other properties of it or to retract the initial attribution. Mark recognition thus imposes a normative order on experience (Kant, KrV, A104). In contemporary jargon, human experience is conceptual or within the space of reasons. But this capacity is also crucial from an inter-subjective angle. Suppose the ox's past leads it associate the stall with distress. It is only insofar as I can represent the fact that one property grounds another, that I can represent the distinction between such property implications and the type of associationistic link manifest in the ox. And it is only insofar as I can draw that distinction that I can be aware of the difference between links that are merely an artefact of my own psychological history (like the ox's fear), and those which, since putatively grounded in the entity itself, should also hold for other agents. In other words, the capacity to use rules to represent objectivity—the entity as a locus of properties with their own implications relations (Kant, KrV, A197/B242)—allows me in turn to represent inter-subjectivity, to represent something as a fact that should hold for a consciousness "in general, not only my own" (Kant, Refl.16, p. 633).[11]

Much of Kant's theoretical philosophy is devoted to unpacking the conditions and the consequences of these ideas and of the capacity for "apperception", which makes them possible. But for current purposes only one more point is needed: these abilities will not typically involve thematic or explicit reflection. To adapt a famous post-Kantian example, when you ask me to pass you a heavier hammer, I unthinkingly reach for the smaller metal one, avoiding the large wooden one: my system automatically represents both the connections between the various materials and weight, and the assumption that those connections, since grounded in those entity, will also hold for you. These facts provide the cognitive infrastructure for practices of explicit justification: if asked why I did what I did, I can explain it, but such justification will typically not in fact take place.

Bringing these points together, we now have a handle on the distinctive capacity for self-determination possessed by Kantian rational agents; it is a capacity to represent, and thus to act on, certain relations and certain distinctions. Since this capacity will be widely manifest in behaviour, there is no need to rely on cases of explicit deliberation when cashing self-determination, something Moran and other prominent Kantians are frequently accused of. The task now is to demonstrate how this material can be used to flesh out the idea of an active/passive distinction. Once that is in place I will then show, in §4, how the combined results can be fruitfully applied to thought insertion.

Activity as commitment

As I noted above, a standard strategy is to argue that inserted thoughts lack some sense of activity, be this "agency" (Stephens and Graham, 1994) or "authorship" (Gerrans, 2001). But it is obviously important to be sure we are operating with the right concepts of activity and passivity here. On the passive side, for example, it is widely recognized that we need to distinguish thought insertion from cases of obsessive thoughts—the phobic's incessant fear that the plane will crash—which lack the distinctive phenomenology of intrusion. But what about the relevant notion of activity? After all, it is vital to keep inserted thoughts separate from unbidden or spontaneous thoughts which simply pop in to my mind; in a great many unpathological cases, a "thought comes when 'it' wants, and not when 'I' want" (Nietzsche, 2002, §17). As Mullins and Spence note, such "unbidden thoughts" create a problem for authors, such as Stephens and Graham, who identify an "experience of mentally acting" as the key feature missing from the insertion case (Spence and Mullins, 2003: 296; the cited remark is from Stephens and Graham, 1994: 2). This is why I diverge from Kantian accounts such as Young's which treats active thoughts as those I "generate … and, importantly, I am aware of generating them" (Young, 2006: 828). I think Young is right that this approach has Kantian backing (for example, Kant, KrV, B158-9), but I think, for the reason just given, it is not the best aspect of Kant's work to utilize here.[12]

As Bortolotti and Broome observe a similar problem will also arise for accounts which, whilst incorporating notions like reason-responsiveness, still frame the issue in terms of belief formation. They give the example of two attempts to identify a broadly rationalist notion of activity, "authorship" which inserted thoughts might lack:

> (a) In order to be the author of the belief that he should file for divorce, Patrick needs *to have formed* that belief on the basis of the best evidence available to him.
>
> (b) In order to be the author of the belief that he should file for divorce, Patrick needs *to be able to endorse* that belief on the basis of the best evidence available to him. (Bortolotti and Broome, 2009: 212)

The problem with (a) is that outside of explicit reflection, few beliefs are so acquired—in addition to the unbidden thoughts just discussed, we often form beliefs and only later come to find good reasons for them. So even if inserted thoughts do lack (a), this cannot be what makes them distinctive (Bortolotti and Broome, 2009: 212–213).

What about Bortolotti and Broome's option (b)? They cash "endorsement" in terms of "the capacity for reason giving in deliberation or justification, or, depending of the type of thought, on behavioural manifestability". More specifically, it:

> [C]ulminates with the subject taking responsibility over the reported attitude and making a commitment to it that is likely to be reflected in further reported attitudes and other forms of behaviour. (Bortolotti and Broome, 2009: 213)

It seems to me that the account at the end of §2 shows how a version of this might work; since I am now going to start importing some specifically Kantian claims, I will call this version "commitment" to distinguish it from Bortolotti and Broome's broader notion of "endorsement".[13] For rational agents, as Kant sees it, to believe is to take on certain commitments, to recognize their implications for your other beliefs, and for the resulting web of norms that governs your behaviour. To use Kant's own example, to see something as a body is to bind oneself to the application of various other properties to it, and to the interrelation of those properties putatively holding for other agents. For a belief to be mine in this sense implies a capacity to undertake a process of explicit justification—as Kant puts it, it must be possible for the "I think" to accompany all my thoughts (Kant, KrV, B131)—but this possibility need not be realised. Rather, a belief is mine insofar as it enters in to the relevant forms of rational combination, combinations which, in line with the results of §2, will be behaviourably manifest.

> "I", in this context, expresses the fact that the representations under consideration are bound and reflected from one standpoint, that of the thinker that refers these representations to herself and commits herself to the unity and consistency of the conceptual ordering of these representations. (Longuenesse, 2007: 153)

This model of activity fits well with Kant's agnosticism (at least in the theoretical domain) as to the true nature of the self: in the absence of knowledge of the noumenal, it is this "relation of rational dependence across mental states that constitutes [our] existence as thinkers" (Kitcher, 2011: 252).

Kant on thought insertion

It is now time to directly apply these materials to the question of thought insertion. The basic Kantian story will include three elements, elements whose positioning differs significantly from much of the literature. First, there is the notion of activity *qua* commitment as defined in §3. Second, there is the notion of ownership: this is to be understood in terms of a willingness to self-ascribe the relevant states. Self-ascription is in turn is understood as Longuenesse suggested: "the thinker that refers these representations to herself and commits herself to [their] unity and consistency" (Longuenesse, 2007: 153). Third, there is some much weaker notion which picks out those representations that are "mine" in the looser, non-Kantian sense—for example, those available to introspection. In Kantian terms, these are accessible via inner sense: a non-rational animal will have representations of its own in this sense, while lacking any that are owned in the stronger senses tied to apperception (Kant, Anth,

127). Defining this third group will be difficult, particularly given the point, made by Kant long before Nietzsche, Freud and others, that the vast majority of what is mine in this extended sense is in fact barely accessible to consciousness (Kant, Anth, 136).[14] What makes Kant's story distinctive, however, is not just the elements in play but their connections. One standard way to handle thought insertion is to frame the discussion in terms of a coming apart of ownership and some relevant notion of activity: inserted thoughts would have the former, but not the latter (for example, Gerrans, 2001). But a Kantian approach will reject this precisely because Kant cashes ownership in terms of commitment: as I noted at the end of the previous section, a thought is mine insofar as it is integrated within the normative web of commitments. As Moran puts it, in this sense "someone determines what shall be part of him as a *person*" (Moran, 2002: 214).

One might have several immediate worries about this proposed taxonomy. On the one hand, it can seem a merely verbal shift—why does it matter if we use "ownership" in a way that is tied to commitment as opposed to mere introspection? Bortolotti and Broome (2009: 216), who propose a similar switch, although on different grounds, provide part of the answer: it does better justice to the phenomenology of inserted thoughts, which are precisely introspected states that the agent nevertheless refuses to accept as her own. But for Kant himself there are also deeper reasons, and these bring me to the other natural objection: far from being merely terminological, doesn't my proposal have the absurd implication that huge numbers of non-pathological states are not really mine? For example, sensations which do not enter into the type of commitment or justification relations sketched: it makes no sense, at both a trivial and a deep level, to say "I think hot", and yet surely I can have a feeling of heat? Or what about deep, recurring desires that I see as central to myself ("how typical of me to want that!") but nevertheless regard as irrational?[15] It is important to see that for Kant this is not a problem, but precisely the result he wants: we need to take very seriously the fact that the "I" for Kant is the "I *think*".[16] This is not the place to enter in to the details of Kantian ethics, but the basic point is that ownership for him is necessary and sufficient for moral responsibility (Kant, GMS, 457). In contrast, sensations and desires (incentives in proper Kantian terminology) are seen as functions of causal chains, stretching back before my birth and over which I have no control, and by extension no direct responsibility (Kant, KrV, A445/ B473, A448/B476).[17] The fact that someone might experience certain sexual desires, for example, is thus conceived as an external force acting on them, as something merely factual. Ownership and responsibility enter the picture only when I take that desire as giving me a reason to act and thus embed it within the web of normative commitments. The result is a sharp divide for Kant between activity, commitment and ownership and freedom on one side, and a vast multitude of representations which are caused to occur "in me", that is, at the point where my body is situated, without thereby being "mine", without being something I own. With respect to these:

> I could not say: "I do it" but must rather say: "I feel in myself an impulse to do it which something has incited in me". (Kant, V-MP/Heinze: 269)

There is clearly a huge amount that is controversial in the views just sketched. But the task now is to apply them to thought insertion. To the lay the groundwork for this, I will take the cases of non-pathological unbidden thoughts, and then obsessive thoughts, before dealing with insertion itself.

First off, unbidden thoughts which are reason-responsive: for example, my sudden realization that I have left the window open. These are unproblematic for Kant for the reasons discussed in §2. They are fully fledged beliefs, through which I thereby take on various commitments (to the conditional that if it rains the floor will be soaked, to the act of shutting it given my other beliefs and desires). The fact that that they arose unbidden is unproblematic since blanket talk of agency has been replaced by notions of commitment and ownership defined in terms of normative integration.

Second, obsessive thoughts, which lack the distinctive pathology of inserted thoughts. This is obviously a huge class ranging from those linked to compulsive disorders ("I must wash my hands again") to images which continually intrude in the mind to an endless posing and re-posing of what most agents would see as settled questions ("do my friends like me" "is the gas on" and so on). Such obsessive thoughts present an interesting case for Kant because they are often not responsive to reasoning: "I can find myself thinking or worrying that *P* even though I realize that there is no reason to think or worry that *P*" (Cassam, 2011: 6). My suggestion is that the Kantian pursue a divide and conquer strategy with respect to the class as a whole. On the one hand, there are cases of systemic bias in reasoning—agents who habitually endorse conclusions that the rest of us regard as unsustainable. This might stem from a particular attitude to the risk calculus, as in the case of agents who systematically over-react to extremely small risks (one can imagine a compulsive handwasher driven by this). Such misguided judgments are, from a Kantian perspective, owned by the agent; she has mistaken views, but they are hers. On the other hand, there are states that are not best seen as actually taking on a commitment (not seen as making a judgment in the technical Kantian sense—Kant, Prol, p.305). Rather they are what O'Brien (2013: 95) nicely calls "mere apprehensions of content". On this model, an image or a thought might insistently intrude into my mind in the way a song or a pain can: standing at the top of the stairs with the baby, an image of it lifeless on the floor flashes before my eyes. For Kant such cases are genuinely passive, genuinely unresponsive to reason and not owned: I cannot say "I think" of them, "I could not say: 'I do it'" but only that external factors are generating a certain state within my body. (Kant, V-MP/Heinze: 269). Whilst she would doubtless not endorse Kant's own use of this type of tactic, O'Brien's own agency approach thus offers a sharp formulation of the basic approach Kant would use for dealing with "non-inserted" but obsessive thoughts:

> In response to the claim that there are 'passive judgements' that are unresponsive to reason the agency theorist can, therefore, insist that either the recalcitrance is consistent with us taking them to be active commitments by the subject to how the world is on some basis, and are thus tractable from the point of view of the agency theory approach to judgement, or are such that it is inconsistent with them being commitments at all. We get the impression that there could be a genuine commitment to the world being thus and so only by thinking it, only by switching between cases which are mere entertainings and cases which are judgements based on unstable reasons. (O'Brien, 2013: 98)

Adding the clarification that such "mere entertainings", whilst well labelled in contrast to judgments, may be radically disruptive and troubling to the subject (for example, consider intrusive images of violence), this is the divide and conquer approach I advocated.

We are now in a position to tackle inserted thoughts directly. I propose the following Kantian analysis. Inserted thoughts are representations that have the semantic and cognitive form of commitments, of "judgments" or "maxims" in Kant's technical

vocabulary—and *yet* which are radically at odds with the subject's existing commitments. The key difference from obsessive thoughts is that obsessive thoughts are either embedded within the agent's normative framework (the first horn of the strategy just sketched), or are "such that it is inconsistent with them being commitments at all" (O'Brien, 2013: 98).

One easy way to approach the proposal is by looking at cases where the inserted thought takes the form of a command:

> Thoughts are put into my mind like 'Kill God'. It's just like my mind working, but it isn't. They come from this chap, Chris. They're his thoughts. (Frith, 1992: 66)

What is happening here is that a representation has entered the system which has normative force—it purports to oblige the subject to do something. In Kantian terms, it is on the active side of the line; it is a judgment or commitment that something ought to be done. The problem is that the individual does not recognize the state's commitments and thus does not see himself as its owner. This is because of its discontinuity with what Longuenesse aptly called "the unity and consistency of the conceptual ordering" of his existing "standpoint". In this sense, the inserted thought is on the passive side of the line: it is not part of the network of self-legislated rules and obligations that define the Kantian I. The result is twofold. First, the subject disavows ownership due to this perceived clash: she cannot say "I think". Second, this disavowal generates the distinctive phenomenology well-captured by Jaspers:

> The patient does not know why he has this thought nor did he not know why he has this thought nor did he intend to have it. He does not feel master of his own thoughts and in addition he feels in the power of some incomprehensible external force. (Jaspers, 1963: 122–123)

This is exactly what Kant would predict: the phenomenology is of being bound by a rule that you did not make yourself. What about cases that lack a straightforward command form? Consider a subject who reports experiencing misogynistic inserted thoughts, a constantly intrusive insistence that "women are *y*" where "*y*" is some string of sexual slurs. From a Kantian perspective, in line with §2, such claims still contain, albeit less obviously, the same command and obligation structure. To judge that *x* is *y*, as opposed to merely associating the two such that the presence of one triggers the presence of the other before the mind, is to represent the former property as grounding the latter, to oblige oneself to attribute the latter to anything that has the former, and to represent the fact that since this attribution is so grounded it should hold for all similar agents and not just for me. Thus to have this judgment in the system is to take on a whole web of commitments as sketched in §2. What patients are reacting to in thought insertion is the fact that something of this form has suddenly been added to their system, a line of normative coding, which they cannot reconcile with what else was already there.

It is important to stress how this differs from mere inconsistency. Doubtless we all have multiple conflicts between our various commitments. But these are typically not manifest to us, and when they become so we either try to repair them, or simply ignore the clash until our attention shifts elsewhere. But in the thought assertion case, a radical clash is presented with an insistent degree of phenomenological saliency: the inserted thoughts don't simply fade in to the background in the way the conflicts between someone's highly paid job and their professed politics might as other things dominate their attention. The result is effectively a persistent avowal of both *P* and not *P*; unsurprisingly, individuals faced with this come to disown one of the thoughts.

The irony of the Kantian framework is this. My strongest desire is, as discussed above, in an important sense not mine; it is something that happens "in me". This is, very crudely, because it is a product of causal forces, ultimately stretching back beyond my control. The "I", in contrast, is defined in normative terms, in terms of the ability to subject my representations and so my behaviour to rules (Kant, KrV, A104-5). *Inserted thoughts have this normative structure, and yet precisely because of it are disowned by agents*. I have focussed on what §2 called the "negative" aspect of thought insertion, the denial of ownership, rather than on the "positive" aspect, the belief that the thoughts belong to some (possibly quite specific) other person. But one can see how the approach sketched might naturally handle that. From a Kantian perspective, a commitment, a judgment, is an act of some agent, an act of binding themselves in various ways.[18] If I am introspectively aware of such a representation and yet am sure that I have not made the relevant commitments, it becomes natural to assume that someone else has done so.

Suppose an agent is faced with this state—what might he or she do? Clearly, one option is to try to remove the inserted thought from the system or at least lessen its prominence; for example, through some form of drug treatment. But from a Kantian perspective, there is another, interesting alternative: try to restore "the unity and consistency of the conceptual ordering" (Longuenesse, 2007: 153). If the divide between the rogue code and their existing perspective is too great, that may be impossible and some form of breakdown will result. But what is striking is that at least in some case reflective patients report trying exactly what the Kantian approach would predict: removing the feeling of alienation by integrating the problematic thoughts so that they become "mine" and not just "in me". This is an exchange between an interviewer, Ira Glass, and Patricia Deegan, a psychologist who suffers from auditory hallucinations:

> IG: How do you conceive of the voices that you hear? As *separate* from yourself, or do you conceive of them as *part of* your self that you can recognize?
>
> PD: I think that for me it's a goal to eventually say these voices are a part of me, and that's actually one of the self-help coping strategies that I do use sometimes...So, for instance, if I have a particularly derogatory or awful voice, that I might say, as a coping strategy, 'today *I am feeling like* I am no good, today *I am feeling like* I'm a worthless person, these are *my* thoughts, these are *my* feelings. (Jenkins and Barrett, 2004: 37)

Conclusion

One of the classic questions Kantian philosophies face is the sustainability of the different dualisms on which they are built. Thus post-Kantian thought often focussed precisely on those elements—such as socially cultivated desires (Hegel) and the transcendental imagination (Hegel and Heidegger)—which might problematize the boundary between active and passive, between spontaneity and sensibility. The case of thought insertion raises similar issues. In a more modern idiom, the position I have outlined is one on which inserted thoughts are unusual precisely because they are both within and without the space of reasons. Obsessive thoughts, in contrast, are easier to regiment as either in or out. Leaving aside these kind of broader issues, it further seems that Kant's account has some interesting implications at the micro-level. One is the link just noted to particular treatment strategies. Another concerns the relationship between self-ascription and activity. For example, Bortolotti and Broome

(2009: 219) identify thought insertion as a failure of both self-ascription and authorship. But, given their definitions, self-ascription without authorship is extremely common (Bortolotti and Broome, 2009: 212): if this is the case, why is the latter's absence sometimes bound up with the radical refusal to self-ascribe seen in thought insertion? On the Kantian picture, what makes inserted thoughts distinctive is not a straightforward failure of commitment, something that is extremely widespread for Kant, but the distinctive *hybrid* state I have discussed.

One way to develop these results would be to use the "inserted emotion" case to drive the wedge in from another angle, so to speak. Consider this report from Mellor:

> I cry, tears roll down my cheeks and I look unhappy, but I have a cold anger because they're using me in this way, and it's not me who's unhappy, but they're projecting unhappiness onto my brain. They project upon me laughter, for no reason, and you have no idea how terrible it is to laugh and look happy and know it's not you, but their emotions. (Mellor, 1970: 17)

The challenge for Kant here is to explain, in effect, how one emotion can be any more heteronomous than another; but that is a task for another paper.

Notes

1 Despite Kant's enormous influence on modern conceptions of the self, there has been relatively little direct research applying his views to the thought insertion case. The only existing items of literature of which I am aware are Chadwick, 1994 and Young, 2006. Whilst I am indebted to both authors, the approach defended here differs substantially, in part because it draws on the last decade's research on Kant's philosophy of mind. I flag particular points of divergence from these accounts as they occur.

2 As is standard, all references to Kant are to the *Akademie* edition, Kant 1902-, using the following abbreviations: Anth: *Anthropologie …* (Ak.7); KrV: *Kritik der reinen Vernunft* (Ak.4); Prol: *Prolegomena* (Ak.4); Log: *Logik* (Ak.9); Refl.:*Reflexion* (Ak.16); SvF: *Die falsche Spitzfindigkeit …* (Ak.2); V-MP/Heinze *Kant Metaphysik L1 (Heinze)* (Ak.28).

3 For a helpful survey see Roessler, 2013.

4 For simplicity's sake, I follow the standard reading of Kant and treat the perceptual states of adult human agents as having the same kind of representational content, namely intuitive judgments, as found in fully meaningful beliefs more generally; in other words, I am leaving aside here both the longstanding issue of judgments lacking intuitive content, and the more recently pressed one of relationalist readings of Kant on perception (see, for example, Gomes, 2014).

5 Perhaps the main omission will be a detailed discussion of the Paralogisms—this is, unfortunately, simply not possible without introducing material regarding Kant's idealism, his relationship to rationalism, and the complex epistemic status of Kantian regulative (as opposed to constitutive) claims about the self, material that would take us too far from the main focus of the article.

6 In addition to pure or transcendental apperception, Kant also talks occasionally of "empirical apperception"; this is equivalent to inner sense (Kant, KrV, A107). In what follows "apperception" will always mean pure or transcendental apperception.

7 Moran's account should equally apply to desires insofar as these are reason responsive—faced with the question of whether I ought to take a job in another country, I reflect on the evidence, deliberate and end by wanting to do so. But the issue of desires is complex in a Kantian context due to the distinction between maxims and incentives—I discuss this below.

8 For an illuminating and sophisticated treatment of the worry see McGeer, 2007: 10.

9 See, for example, Crowell, 2007.

10 Such inter-property implications may be analytic or synthetic.

11 I might be wrong in any particular case—we mistake mere associations for facts all the time. But what is significant for Kant is the ability to represent the distinction (similarly, the Second Analogy is an attempt to show that the ability to draw a distinction between a succession of perceptions and a perception of succession depends on a capacity to deploy the categories; this is perfectly compatible with the fact that we may be wrong about which is which in a given case).

12 Generally, it seems that the cases of automatic or fluid action emphasized by post-Kantian phenomenology should make us wonder how often an experience of "generating" thoughts is ever really present. For a classic formulation of the phenomenological critique, see Dreyfus, 1991. One other point of difference between Young and I is that he assigns a central role to Kant's complex theory of the imagination; I touch on this at the end, but my position in no way relies on it.

13 In virtue of its explicitly Kantian heritage, the account I am outlining differs from Bortolotti and Broome on several points. For example, I am—in virtue of the facts about Kantian semantics noted in footnote 4—more willing to apply it to perceptual beliefs than they are (Bortolotti and Broome, 2009: 210). I am also unclear how strongly they intend the requirement regarding reason giving: what, for example, of beliefs which the subject has inherited unquestioned? (Bortolotti and Broome, 2009: 212–213). On my account such beliefs remain on the active side; what matters is the commitments an agent takes on, even if she can offer no more illuminating justification than "that is just how things are". I return to another important point of difference from Bortolotti and Broome below; despite these differences, I am indebted to their work.

14 In opposition to my approach, Young aligns this loose grouping with the "I think" (Young, 2006: 830). There are many ways one might regiment the terminology, but it seems to me Young's tactic of aligning apperception with what he calls "global" representation will create problems with regards to animals, since Kant is insistent these lack the "I think" (Kant, Anth, p.127). At a primary text level, the issue is visible in Kant's vacillations over whether to allow animals "consciousness" [*Bewustßein*] (for a close study of the relevant passage see McLear, 2011).

15 In Kantian technical terms, incentives or *Triebfedern* that are not taken up into maxims. I am supressing complications here concerning the precise way in which incentives might be incorporated into maxims and potentially harmonized with our other commitments and with obligations such as those set out by the categorical imperative; their recognition would not affect the basic point of the paragraph, but would require extensive treatment of issues surrounding the so called "Incorporation Principle" (the locus classicus here is Allison, 1990).

16 Or the "I do", that is, "the I endorse this maxim"—see the citation at the end of the paragraph.

17 I say "likely" as I cannot deal here with the complex issue of our responsibility not to enter situations in which, the relevant causal chains being what they are, my incentives will tend further in directions that I ought not to endorse (Kant, GMS, 399).

18 I am supressing various complexities regarding the relation between spontaneity, which I take to be exercised in taking on conceptual commitments in this fashion, and full blown Kantian autonomy which involves self-legislation in an even stronger sense (crudely, I not only give myself some law like "treat all x as y" I, or rather my rational nature, is the full source of the law's content). Treatment of this issue would require a discussion of the theoretical/practical distinction in Kant which is beyond the scope of this article.

References

Allison H (1990) *Kant's Theory of Freedom*. Cambridge University Press: Cambridge, UK.

Andreasen NC (1984) *Scale for the Assessment of Positive Symptons*. University of Iowa: Iowa.

Bortolotti L and Broome M (2009) A role for ownership and authorship in the analysis of thought insertion. *Phenomenology and the Cognitive Sciences*; **8** (2): 205–224.

Boyle M (2011) Making up your mind' and the activity of reason. *Philosophers' Imprint*; **11** (17): 1–24.

Cassam (2011) Knowing what I believe. *Proceedings of the Aristotelian Society CXI (I)*; 1–23.

Chadwick R (1994) Kant, thought insertion and mental unity. *Philosophy, Psychiatry and Psychology*; **1** (2): 105–113.

Crowell S (2007) Sorge or Selbstbewußtsein? Heidegger and Korsgaard on the Sources of Normativity. *European Journal of Philosophy*; **15** (3): 315–333.

Dreyfus H (1991) *Being-in-the-World*. MIT Press: Cambridge, MA.

Evans G (1982) *The Varieties of Reference*. Clarendon Press: Oxford.

Frith CD (1992) *The Cognitive Neuropsychology of Schizophrenia*. Lawrence Erlbaum Associates: Hove, UK.

Fulford KWM (1989) *Moral Theory and Medical Practice*. Cambridge University Press: Cambridge, UK.

Gerrans P (2001) Authorship and ownership of thoughts. *Philosophy, Psychiatry and Psychology*; **8** (2/3): 231–237.

Gomes A (2014) Kant on perception. *Philosophical Quarterly*; **64** (254): 1–25.

Golob S (2016) Self-Knowledge, transparency and self-authorship. *Proceedings of the Aristotelian Society CXV* (3): 235–253.

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

Jaspers K (1963) *General Psychopathology*. Manchester University Press: Manchester, UK.

Jenkins JD and Barrett RJ (2004) *Schizophrenia, Culture, and Subjectivity*. Cambridge University Press: Cambridge, UK.

Kitcher P (2011) *Kant's Thinker*. Oxford University Press: Oxford.

Longuenesse B (2007) Kant on the identity of persons. *Proceedings of the Aristotelian Society*; **107** (1): 149–167.

McGeer V (2007) The moral development of first-person authority. *European Journal of Philosophy*; **16** (1): 81–108.
McLear (2011) Kant on animal consciousness. *Philosopher's Imprint*; **11** (15): 1–16.
Mellor CS (1970) First rank symptoms of Schizophrenia. *British Journal of Psychiatry*; **117** (536): 15–23.
Moran (2001) *Authority and Estrangement*. Princeton University Press: Princeton, NJ.
Moran (2002) Frankfurt on Identification. In: Buss and Overton (eds.) *Contours of Agency*. MIT Press: London, pp 188–217.
Moran R (2004) Precis of authority and estrangement. *Philosophy and Phenomenological Research*; **69** (3): 423–426.
Nietzsche F (2002) *Beyond Good and Evil*. Cambridge University Press: Cambridge, UK.
O'Brien L (2013) Obsessive thoughts and inner voices. *Philosophical Issues*; **23**, 93–108.
Roessler J (2013) Thought insertion, self-awareness, and rationality. In: Thornton T, Graham G, Sadler J and Davies M (eds.) *The Oxford Handbook of Philosophy and Psychiatry*. Oxford University Press: Oxford, pp 658–672.
Shah N and Velleman D (2005) Doxastic deliberation. *Philosophical Review*; **114** (4): 497–534.
Spence S and Mullins S (2003) Re-examining thought insertion. *British Journal of Psychiatry*; **182**, 293–298.
Stephens GL and Graham G (1994) Self-consciousness, mental agency and the clinical psychopathology of thought insertion. *Philosophy, Psychiatry and Psychology*; **1** (1): 1–10.
Wing JK, Cooper JE *et al* (1974) *Measurement and Classification of Psychiatric Symptoms*. Cambridge University Press: Cambridge, UK.
Young G (2006) Kant and the phenomenon of inserted thoughts. *Philosophical Psychology*; **19** (6): 823–837.

Data availability

Data sharing not applicable to this article as datasets were neither generated nor analysed.

Additional information

Competing interests: The Authors declare no competing financial interests.

Reprints and permission information is available at http://www.palgrave-journals.com/pal/authors/rights_and_permissions.html

How to cite this article: Golob S (2017) Kant and thought insertion. *Palgrave Communications*. 3:16108 doi: 10.1057/palcomms.2016.108.