teorema
Vol. XXXIV/1, 2015, pp. 149-160
ISSN: 0210-1602
[BIBLID 0210-1602 (2015) 34:1; pp. 149-160]

Replies to My Critics

Jordi Fernández

First of all, I want to thank Josep Lluís Prades, Lisa Bortolotti, Kengo Miyazono and André Gallois for their insightful commentaries on *Transparent Minds*. I am not sure that I will be able to do full justice to their commentaries in my replies, but I hope that I have been able to concentrate on their most substantive points regarding the book. I am also grateful to the editors of *Teorema* for giving us the opportunity to discuss *Transparent Minds* in this journal.

1. *Prades on Our Grounds for Belief and Desire*

Josep Lluís Prades raises two worries about the bypass model, both of which ultimately concern the use that I make of the notion of grounds. Prades's first worry is that the conditions that the model requires for self-knowledge of belief are too demanding. For there seem to be cases in which, intuitively enough, a subject knows the belief that she is holding even though there is no such thing as the subject's grounds for that belief. If I look at a stick in a glass of water, Prades argues, then the stick has the visual appearance of being slightly bent even though not only do I believe that it is a completely straight stick, but I also know that I believe it. What explains the fact that I know that I have that belief? It cannot be the fact that I self-attribute the first-order belief on the basis of my grounds for believing that the stick is straight, Prades claims, since the visual appearance of the stick constitutes grounds for believing that the stick is bent, and not straight. One could protest that, by itself, the visual appearance of the stick does not qualify as my grounds for believing that the stick is straight. Strictly speaking, my grounds for that belief are constituted by a combination of the perceptual state in which I am when I look at

the stick, plus the belief that I am experiencing a perceptual illusion. Prades anticipates this move and, in reply, he puts forward the following variation of the original case. Suppose that I momentarily experience some anomalous psychological condition and, as a result, I briefly trust the visual appearance of the stick, so I form the belief that the stick that I am seeing is bent. Once again, it intuitively seems that I know that I believe that the stick is bent, and yet I have no grounds for believing that the stick is bent. For the type of perceptual experience that I am having when I look at the stick in the glass of water does not tend to produce in me the belief that the stick that I am seeing is bent. It does so on this occasion but, Prades tells us, these are abnormal circumstances. It so happened that the perceptual experience caused me to believe that the stick is bent, but it would not normally do that. More generally, the difficulty for the bypass model is, according to Prades, that in a situation in which the formation of the first-order belief is sufficiently unusual for me not to have grounds for that belief, it nonetheless seems that I have no trouble knowing that I am having the belief in question. Prades believes that the bypass model has analogous difficulties explaining our knowledge of our own ungrounded desires.

It is hard to address this type of case without knowing what unusual psychological condition I am supposed to be in when, in Prades's case, I look at the stick in the glass of water. I agree with Prades that the bypass model predicts that if I form some first-order belief by a total fluke, then I do not qualify as knowing that I have that belief. As I see it, though, this is not such a counter-intuitive prediction. If the psychological circumstances that I am experiencing when I form the fluky first-order belief are quite abnormal (and they will need to be seriously abnormal for me to lack any sort of grounds for my first-order belief), then it does not seem unreasonable to think that, even though my self-attribution of the first-order belief is true, it is not justified. After all, if I have taken drugs, or I have not slept in the last 48 hours, and I am not responsible for the first-order beliefs that I am forming, then what reason is there to think that I will be competent in attributing beliefs to myself? Similar considerations apply to our alleged self-knowledge of ungrounded desires. If my state of mind is disturbed enough for me to want things which I regard as worthless, useless and unappealing, then why should we assume that it is within my cognitive power to attribute desires to myself proficiently?

Prades's second worry about the bypass model is that the explanation it provides of self-knowledge for desire makes some assumptions about the nature of desire which are wrong. Specifically, it makes two wrong assumptions. The model assumes, first of all, that urges and values can play a grounding role for desire. Urges and values, Prades claims, are just desires by another name. He points out that, given some physiological imbalance, I may need to swallow substances that I find repulsive. It is true that, in this kind of case, I have neither the desire to swallow the relevant substance, nor the urge to do it. But surely that does not show that the desire and the urge are identical. The reason for differentiating urges (appetites, cravings, yearnings, longings) from desires is that it is possible to describe situations in which a subject has an urge for something without desiring it. The subject may have reasons for restraining herself from forming the relevant desire, or she may lack the necessary concepts to frame it. An analogous point applies to values. A subject may find that a fulfilling romantic relationship, for example, would be a good thing for her to have, but she may not be at all inclined to pursue one because, let us say, she is deeply depressed. It seems natural to describe that subject by saying that she values a fulfilling romantic relationship, but she does not want one.

Secondly, Prades tells us, the bypass model wrongly assumes that a subject's desire for a certain end, combined with the belief in certain means to that end, provides the subject with grounds for a desire to pursue those means. After all, my desire to end a headache and my belief that if someone cuts off my head, then the pain will stop does not make me want to have my head cut off. I do not disagree. The 'production of desire' principle to which Prades refers is introduced as a tendency law precisely to accommodate cases of this kind. There are various reasons why a subject may not want to pursue those things which, from the point of view of her own beliefs, would lead to the satisfaction of one of her desires. Weakness of will may be one of those reasons. The conflict between satisfying the relevant desire and satisfying the subject's desire for survival may be another. This is, in my view, what Prades's headache case illustrates.

Prades is right in concentrating on the notion of grounds in his discussion of the bypass model. For that notion is meant to do a lot of work within the model. As Prades points out, within the bypass model the notion of grounds is a causal, and not a normative, notion. This is

deliberate. A purely causal notion of grounds allows the model to explain, for example, why the smoker who wants a cigarette knows that she does despite not finding smoking desirable. But this benefit comes at a cost, as Prades's cases illustrate. I acknowledge that the transparency theorist needs to say more about the relation that needs to hold between a mental state and a subject's belief, for example, in order for that mental state to qualify as the subject's grounds for her belief. I am inclined to think that a functional characterization of the belief should give us an answer to this question. But this is a line of enquiry that I cannot pursue here.

2. *Gallois on the Redundancy of the Bypass Model*

André Gallois's thorough discussion of the bypass model raises four worries for the model. The first one is that the notion of justification employed by the model is inadequate. The second one is that the model's conditions for self-knowledge are too undemanding. The third one is that the model does not explain why we are more justified in self-attributing beliefs than other people are in attributing those beliefs to us. The fourth one is that the model only provides a redundant solution to the three philosophical problems to which it is applied. Let us take these worries in order.

Gallois claims that the bypass model relies on a straightforward regularity account of justification. He also thinks that there are counterexamples to such accounts of justification. I agree on the latter point. Suppose, for example, that I have a thermostat-like mechanism implanted in my brain which causes me to believe that the temperature around me is 20C if and only if it is indeed 20C. We still wouldn't want to say that, whenever the mechanism implanted in my brain triggers the belief that the temperature is 20C, I am justified in believing that. But Gallois's description of the notion of justification employed by the bypass model is not completely accurate. According to the relevant notion of justification, a subject is justified in forming a belief if she forms it on the basis of a state that constitutes adequate support for it. Now, it is true that all it takes for a state to constitute adequate support for a belief is a certain regularity; the regularity between the occurrence of that state and the truth of the belief. However, in order for the subject to form her belief on the basis of that state, that state needs to be readily available to the subject. In what sense? She needs to be

disposed to believe that she is in that state if, for example, her belief is challenged and she is asked to produce reasons in support of it. This condition is meant to rule out counterexamples to pure regularity accounts of justification such as the counter-example sketched above.

Gallois also thinks that the conditions that the model requires for self-knowledge of belief are too undemanding. For there seem to be cases in which, intuitively enough, a subject is not justified in forming a certain belief even though the conditions for self-knowledge required by the bypass model are satisfied. Suppose that Samantha finds a highly reputable historical text according to which Constantinople fell in 1453. If the bypass model is correct, Gallois tells us, then Samantha is justified in believing that if the text says that Constantinople fell in 1453, then she believes that it did. Why is that? Because if the bypass model is correct, then the fact that the text says that Constantinople fell in 1453 justifies Samantha in believing that she believes it did. And, in general, if P justifies someone in believing that Q, then she is justified in believing that if P then Q. And yet, Gallois tells us, Samantha is not justified in believing that if the text says Constantinople fell in 1453, then she believes that it did.

Notice, though, that the bypass conditions for self-knowledge do not yield, by themselves, the result that Samantha is justified in believing that if the text says that Constantinople fell in 1453, then she believes that it did. To reach that outcome, Gallois also appeals to the general principle that if P justifies someone in believing that Q, then she is justified in believing that if P then Q. I am inclined to challenge that general principle. There are cases in which someone can be given excellent evidence to believe something but, due to prejudice or bias, she will refuse to accept the evidence, and she knows that she will. Suppose, for example, that I am given excellent biological evidence to believe in natural selection. However, it turns out that, due to my religious beliefs, I will not accept any evidence to believe in natural selection. Moreover, I am aware of having that bias. In that situation, the evidence provided to me justifies me in believing in natural selection, since my bias doesn't affect the quality of the evidence provided to me. And yet, I am not justified in believing that if the evidence is right, then I believe in natural selection. After all, I know full well that nothing could convince me that natural selection is real.

According to Gallois, the bypass model does not explain why a subject is more justified in self-attributing a belief than other people

are in attributing that belief to her. To motivate this worry, Gallois puts forward the following example. Suppose that I am justified in self-attributing the belief that you are in pain. According to the bypass model, my self-attribution is justified by the very same state which may justify my belief that you are in pain, namely, the perceptual state in which I am when I apparently perceive your pain behaviour. Why then think, Gallois asks, that less can go wrong when I self-attribute my belief that you are in pain than when I attribute pain to you? After all, my grounds are the same. The reason is that, even though my grounds are constituted by the very same state (namely, my perception of your pain behaviour), that state justifies my attribution of pain to you and my self-attribution of the belief that you are in pain in virtue of different facts, facts which can come apart. Thus, if my perceptual apparata are unreliable unbeknownst to me, then I am likely to make a mistake in my attribution of pain to you, but that does not make my self-attribution of the belief that you are in pain vulnerable to error. Provided that I continue to take my perceptual experiences at face value, I am not likely to be mistaken in my self-attribution of that belief. That is why my attribution of pain to you is vulnerable to error in ways in which my self-attribution of the belief that you are in pain is not. Which explains, in turn, why the former is less justified than the latter.

The bypass view can be deployed to deliver an account of the thought insertion delusion, a solution to Moore's paradox and an explanation of self-deception. On Gallois's view, however, the bypass model does no real work in illuminating the three phenomena. Let us consider, then, Gallois's concerns regarding the three applications of the bypass model.

Take the thought insertion delusion first. Why does the thought insertion patient believe that she has a certain belief that it is not hers? Because, even though she can attribute the relevant belief to herself, that self-attribution does not put pressure on her to have the relevant belief. In that sense, the self-attribution is not 'assertive.' The bypass model of self-knowledge offers an explanation for why, normally, our self-attributions of beliefs are assertive. If we self-attribute those beliefs on the basis of our grounds for them, then it is no wonder that we are inclined to have the beliefs that we attribute to ourselves. After all, our self-attributions of beliefs should make us recognize that we have grounds for having those beliefs. If this is correct, then the bypass model suggests a reason for why the thought insertion patient believes

that she has a belief that it is not hers. The reason is that the patient has not been able to attribute the belief to herself through bypass. Gallois does not take issue with the first part of this explanation. He is prepared to concede that the thought insertion patient's self-attribution of the 'inserted' belief may not be assertive. But he does not think that the bypass model explains why, normally, our self-attributions of beliefs are assertive. If Gallois is correct, then assertiveness plays a role in explaining the delusion but, since the bypass model does not explain assertiveness, the bypass model plays no role in explaining thought insertion.

Why does Gallois think that the bypass model does not explain assertiveness? Because he thinks that there is no reason why a subject should become aware of her grounds for a first-order belief in the process of self-attributing it. But there is one. The reason is that, as noted above, in order for a subject to form the belief that she has a certain belief on the basis of her grounds for that first-order belief, those grounds need to be readily available to her. I take this to be part of what the basing relationship requires. And if those grounds are readily available to the subject (in virtue of the fact that she self-attributes the relevant first-order belief on their basis), then, in the scenario in which the question of whether she has the first-order belief arises, she should be aware of her grounds for having that belief. And this, in turn, should put pressure on her to have the belief, which explains why our self-attributions of beliefs are assertive.

Gallois has a similar concern regarding the solution to Moore's paradox offered by the bypass model. Suppose that Samantha believes the following conjunction: Moore was a philosopher, and I do not believe that Moore was a philosopher. Why is she being irrational? The bypass model tells us that if she believes that she does not believe that Moore was a philosopher, it is because she finds no grounds for believing that Moore was a philosopher. But if she finds no grounds for believing that Moore was a philosopher, then surely it is irrational for her to believe that Moore was a philosopher. And yet, she does believe it. In reply, Gallois suggests that, when Samantha believes that she does not believe that Moore was a philosopher, she may fail to recognize that she has a reason to believe that Moore was a philosopher. I, however, have trouble seeing how Samantha could have formed, then, the belief that Moore was a philosopher on the basis of that reason. How can a single mental state be, on the one hand, available enough to

Samantha for her to form the belief that Moore was a philosopher on the basis of that state while, on the other hand, it is not available to her when she wonders whether she believes that Moore was a philosopher?

Gallois is not persuaded by the account of self-deception offered by the bypass model either. According to this account, if we find Jack blameworthy for behaving in a way that indicates that he believes that he is sick while, at the same time, Jack denies that he has that belief, it is because we sense that Jack has committed a certain form of epistemic negligence. If the bypass model is correct, then Jack has formed his belief that he does not believe that he is sick upon finding no grounds for the belief that he is sick. But Jack's behaviour suggests that he thinks that he is sick. Thus, Jack has formed a certain belief despite finding no grounds for it. And it should be evident to Jack that he has found no grounds for it, since that fact is precisely what supports his higher-order belief. Now, Gallois thinks that the bypass model is not necessary to account for self-deception in terms of epistemic negligence. Why is it not equally plausible for an introspectionist, Gallois asks, to claim that Jack is at fault in forming his higher-order belief because he fails to detect both his first-order belief and his grounds for holding it?

I am not sure that making a mistake in introspecting one's own beliefs needs to amount to negligence. Is the introspecting subject's failure to detect his first-order belief something that he can be blamed for? Presumably, this depends on the reasons why he fails to detect his first-order belief. By contrast, the bypass model guarantees that the self-deceived subject can be blamed for his failing to obtain self-knowledge. For it yields the result that Jack, for instance, is disregarding his lack of grounds for forming beliefs, which is a form of negligence. Admittedly, there may be reasons why an error in the process of acquiring self-knowledge can amount to negligence other than those highlighted by the bypass model. So I am not claiming that negligence can only be invoked in an explanation of self-deception through the bypass model. I only claim that the bypass model delivers one form of epistemic negligence in the cases of self-deception concerned. The way Gallois sees it, this means that the bypass model is redundant: All the work in the explanation of self-deception is done by the notion of epistemic negligence, and none by the bypass model. The way I see it, the bypass model does the work of ensuring that epistemic negligence has taken place in self-deception.

3. *Bortolotti and Miyazono on Why Inserted Thoughts are not Beliefs*

In their commentary, Lisa Bortolotti and Kengo Miyazono focus on the account of the thought insertion delusion based on the bypass model of self-knowledge. We have seen above that, according to this account, the thought insertion patient thinks that she has a certain belief, but this does not put pressure on her to have the relevant belief. What the bypass model of self-knowledge provides is an explanation of where the relevant pressure comes from in the normal case. In the normal case, the explanation goes, we self-attribute our beliefs on the basis of our grounds for them, which makes our grounds for having those beliefs salient to us. And this, in turn, puts pressure on us to have those beliefs. If this is right, then the reason why the thought insertion patient believes that she has a certain belief which is not hers is that the patient has not been able to attribute the belief to herself through bypass.

Bortolotti and Miyazono agree that the thought insertion patient does not endorse the 'inserted thought', or (I take this to be equivalent) she is not committed to the truth of it. What Bortolotti and Miyazono take issue with is my assumption that the 'inserted thought' is a belief. This is a substantive challenge. For if the thought insertion patient is not referring to a belief when she claims to have a thought that is not hers, then, on the face of it, the bypass model does not explain why the patient lacks the phenomenology of feeling pressured to have that mental state. Let us consider, then, Bortolotti and Miyazono's reasons for rejecting the idea that 'inserted thoughts' are beliefs.

The first reason is that thought insertion patients never mention beliefs in their reports. This is, to the best of my knowledge, correct. I am not sure, however, that we can expect patients to report their beliefs by using the term 'belief.' Most of us would refer to our beliefs as opinions, beliefs, notions, ideas or thoughts indistinctly, unless we were familiar with the relevant philosophical distinctions. So I assume that thought insertion patients are not different from other people who are unfamiliar with the technical notions of belief and thought, and they use terms such as 'thought' and 'idea' to refer to their beliefs. I acknowledge that proceeding thus is not taking their reports at face value. Neither is, for that matter, to claim that thought insertion patients do not experience being the agents of their thoughts. As far as I am aware, every account of the thought insertion delusion on offer re-

interprets, somewhat creatively, some bits of the patients' reports. In my opinion, this is a cost that every theorist needs to pay for an account of the delusion. It is the cost of making sense of reports which would be unintelligible if they were entirely taken at face value. But the cost does need to be kept at a minimum, and it does need to be disclosed in the presentation of one's account.

The second, and more important, reason is that the consideration I offer in favor of interpreting the patients' talk about 'thoughts' as talk about beliefs is insufficient. In my view, so-called inserted thoughts cannot be thoughts because thoughts come to our minds unsolicited all the time, and we do not find that puzzling. By contrast, thought insertion patients must find something very odd in their experiences of so-called inserted thoughts. Why is that? Because they claim that those mental states are not theirs. This claim, I take it, is the expression of a deeply abnormal experience. If the mental states at issue are beliefs, and the patients do not feel committed to their truth, then we can see why they would be puzzled at the fact that they seem to be having them. By contrast, if those mental states consist in merely entertaining certain propositions (such as 'the garden looks nice', 'I am especially bad' or 'I should kill Lissi'), then it is hard to see why thought insertion patients would be puzzled when they experience having them.

Or is it? Bortolotti and Miyazono do not think it is hard to see at all. The patients' puzzlement, they propose, can be easily explained by the fact that they find, in their minds, some thoughts that are not owned. What is puzzling for those patients, Bortolotti and Miyazono tell us, is that a thought that is not owned is present in the patient's mind. It is not clear to me what notion of ownership Bortolotti and Miyazono are employing exactly. My initial impression was that, according to them, ownership of a mental state consisted in self-attributing that mental state; believing that the state is one's own. But it seems to me that if that is the notion of ownership at play, then the alternative explanation proposed by Bortolotti and Miyazono turns out to be vacuous.

Let us keep in mind that what needs to be explained is the puzzlement that leads thought insertion patients to claim that the 'inserted' mental states are not theirs. Surely the explanation of why experiencing those mental states is puzzling for them cannot be that those patients disown the relevant mental states, since disowning a mental state and thinking that the mental state is not one's own is, on the notion of ownership concerned, one and the same thing. What we want

to explain, in the first place, is why those patients are puzzled at finding certain mental states in their own minds; puzzled to the point of claiming that those mental states are not theirs. That is our *explanandum*. So how can the fact that those patients disown (that is, fail to self-attribute) the relevant mental states be part of our *explanans*?

Perhaps what Bortolotti and Miyazono have in mind is that ownership of a mental state consists in endorsing the mental state at issue, or being committed to the truth of it. In that case, it makes sense to claim that what explains the puzzlement of thought insertion patients at finding certain mental states in their minds, mental states which they claim are not theirs, is the fact that those mental states are not owned by the patients. For in that case, what is meant by 'not owned' is that the patients do not endorse the truth of those mental states. On this notion of ownership, then, the claim 'the puzzlement of thought insertion patients at finding certain mental states in their minds, mental states which they claim are not theirs, is explained by the fact that those mental states are not owned by the patients' does not turn out to be circular. But the claim is, as far as I can see, implausible if the relevant mental states are assumed to be thoughts. I have trouble seeing why the thought insertion patient would be puzzled at finding a thought that is not owned (that is, not endorsed) in her mind. As Bortolotti and Miyazono point out, it is not strange for us to find episodes of imagination, whose truth we are not committed to, in our own minds. So why would thoughts be any different in that respect?

A separate concern raised by Bortolotti and Miyazono has to do with two characteristic features of delusions; their little weight in informing action, and their resistance to counter-evidence. In support of the account of the thought insertion delusion provided by the bypass model, I offer, in *Transparent Minds*, the consideration that if the account is right, then it sheds light on why the thought insertion delusion is resistant to counter-evidence, and why it has little weight in informing action. In response, Bortolotti and Miyazono put forward evidence that delusions of thought insertion do make a difference to the subject's behaviour. For instance, a thought insertion patient can pursue certain actions that are aimed at preventing the insertion of alien thoughts. This is a good point. However, it seems to me that my view is not inconsistent with it.

My view is that, since thought insertion patients do not feel that their being aware of their 'inserted thoughts' puts pressure on them to

endorse the truth of those mental states, it is not surprising that those patients are not inclined to perform a number of actions that, in the normal case, we would feel inclined to perform if we were aware of having those mental states. Which are those actions? Assuming that the relevant mental states are beliefs, the actions at issue are those actions which we would be inclined to perform if we endorsed the truth of those mental states upon finding them in our own minds. If I thought, for example, that I believed that I should murder Lissi, then I would be inclined to plan Lissi's murder because my self-attribution of that belief would make me endorse the proposition that I should murder Lissi. If I thought that I believed that the garden looks nice, then I would be inclined to, let us say, take a photograph of it, or perhaps show my garden to other people, because my self-attribution of that belief would make me endorse the proposition that the garden looks nice. But those are actions that we cannot expect the relevant thought insertion patients to be inclined to perform when they find that they have 'inserted thoughts' with those contents, since their self-attributions of their 'inserted thoughts' put no pressure on them to endorse the truth of those mental states.

All this seems to be consistent with the fact that there may be *other* actions which the thought insertion patients are inclined to perform in virtue of having their delusion; actions which do not concern whether the 'inserted thoughts' are correct or not. To be precise, therefore, my view is not that the thought insertion patient is not inclined to act, in any way, upon her delusion that she has a certain thought. It is only that she is not inclined to act upon her delusion in a number of ways in which we would be inclined to behave if we thought that we had the same thought. I can see, however, why Bortolotti and Miyazono might complain that abbreviating this view by saying that the thought insertion delusion has 'little weight' in informing action is misleading.

School of Humanities
University of Adelaide
Adelaide SA 5005, Australia
E-mail: jorge.fernandez@adelaide.edu.au