Issue 5(1) Summer Conference 2012

(Print) ISSN 1748-9393
(Online) ISSN 2051-5359

British Journal of Undergraduate Philosophy

Editor-in-chief: Matthew Green
University of Leeds

Journal of the British Undergraduate Philosophy Society

British Journal of Undergraduate Philosophy

The Journal of the British Undergraduate Philosophy Society
Issue 5(1) Spring 2012

www.bups.org/bjup-online
(Print) ISSN 1748-9393 | (Online) ISSN 2051-5359

British Journal of Undergraduate Philosophy

www.bups.org/bjup-online

(Print) ISSN 1748-9393 | (Online) ISSN 2051-5359

Acknowledgements

Thanks to Timothy Williamson for agreeing to delivery the keynote speech at the British Undergraduate Philosophy Society Summer Conference 2012, at which the first five papers in this issue were delivered; to Kit Fine, John McDowell, Ernest Sosa, Barry Stroud, and Timothy Williamson for agreeing to be interviewed for this issue; to Steve Pyke for allowing us to use his photgraphs of famous contemporary philosophers; to our anonymous peer-reviewers for giving us such full feedback on all submitted papers, often at very short notice; to the Association for Symbolic Logic, Cambridge University Press, Professor Graeme Gooday, Logic Colloquium, Mind Association, Society for Women in Philosophy, and The University of Leeds Department of Philosophy for their generous financial support of the conference; and to Jord Adshead for the work done in his former position of commissioning editor, and Jade Fletcher and Alexander Roberts for their help with manuscript editing. I'm especially thankful to Emma Cockeram, Daniel Houston, Emily James, Dino Jakušić, and Michael Lyons for their great work as part of the British Undergraduate Philosophy Society committee, and to Daniel Houston, Dino Jakušić, and Michael Lyons in particular for their great work on this issue of the British Journal of Undergraduate Philosophy. MG.

Contents

Agency, Frankfurt-Cases and the Compatibility of Determinism with Free Will and Moral Responsibility*

Alex Moran
University College London

Abstract

In 1969, Frankfurt attempted to undercut the debate between compatibilists and incompatibilists by arguing that the Principle of Alternate Possibilities (PAP: an agent acts responsibly only if he could have done otherwise) is false. He did this by constructing what are now known as 'Frankfurt-cases'; cases in which he alleged it is both true that an agent could not have acted otherwise and yet also that the agent acted freely and responsibly. This paper discusses a challenge to Frankfurt's interpretation of these cases; one which offers a theory of agency that casts doubt on their conceptual coherence and which can be used to offer a novel argument for incompatibilism. My argument, in brief, is that the challenge is not sufficiently motivated to pose a problem for Frankfurt's interpretation of these cases, or to provide a strong argument for incompatibilism.

Introduction

The main thesis of Frankfurt [13] is that the Principle of Alternate Possibilities (PAP), which states that an agent acts responsibly only if he could have done otherwise, is false. To show this, he asks us to consider the following sort of case. Suppose Jones is going to kill Smith. Suppose also that Black, who wants to make sure Smith gets killed, is monitoring Jones, and has all the requisite powers to detect if he is going to change his mind and to make him kill Smith if this change occurs. Lastly, suppose that Black does not wish to get involved unless he has to. Now, when Jones decides, of his own accord, to kill Smith, and then proceeds to kill him, three things seem to be true: (a) Jones kills Smith, (b) Jones acts freely and responsibly, and (c) if he had not so decided, he would have killed Smith anyway – that is, he could not have done otherwise.1 This case, then, seems to falsify PAP, by showing that an

*Delivered at the BUPS Summer Conference 2012 on 2-3 June 2012 at the University of Leeds.

agent can act freely and responsibly despite being unable to do other than that which he did.

There are various ways to challenge Frankfurt's interpretation of these sorts of cases. The one I wish to focus on starts from considerations of agency in order to challenge their conceptual coherence.[1] This it does by arguing that there is no way to construct a case in which the counterfactual scenario is such that Black will make certain that Jones kills Smith, because if the event 'Smith's death' is made certain in this way, then Jones could not have done otherwise, and if Jones could not have done otherwise, then Jones didn't perform an action at all. This, it is argued, is because an event only qualifies as a genuine action by an agent, if that agent possessed in respect of it the dual power to Φ or not-Φ in such a way that it was exercisable 'at the time of the action' (henceforth 'at t').[2] If this is right, then the compatibilist (i.e. Frankfurt's) interpretation of the above sort of case does seem problematic. For the two essential features of a Frankfurt-case are that (1) Jones performs an action for which he is morally responsible, and (2) Jones could not have avoided performing that action. But if the above condition on agency is true, then (1) and (2) can never be met at the same time, since (2) entails the falsity of (1).

What exactly it means to be able to do other than one does will be discussed in more detail below. For now, it will be sufficient to note that it does not require the capacity to be bring about, as Steward puts it, another 'positive chain of events'[3]; but only that one be able to refrain from performing that action. Steward, in fact, sees this argument as offering a replacement principle for PAP. But I think it is better understood – and this seems to be how Alvarez understands it – as a clarification of it. For we can view the argument as suggesting that the following condition on agency holds:

Agency Condition (AC): An event only qualifies as a genuine action by an agent if she could have refrained from performing it (viz. the dual power to Φ or not-Φ which was exercisable at t);

and that if AC holds, then the following principle also holds:

Principle of Possible Refrainment (PPR): An agent acts freely and responsibly only if she could have refrained from so acting.

[1] The argument is due to Steward [23], [24] and Alvarez [1].

[2] Steward [23] p. 24.

[3] Ibid. p. 77.

PPR is clearly a clarification of PAP; a clarification which switches the emphasis from the agent's power to do other than she does, to her power to refrain from doing that which she does. However, PPR does differ significantly from the traditional interpretation of PAP. This is because it does not treat PAP as a principle which isolates a necessary condition on moral responsibility, as has been traditionally the case. Instead, it treats PAP as highlighting a necessary condition on agency, and therefore, (derivatively) on freedom and moral responsibility. PPR, then, may be used not only to rebut Frankfurt-type-cases, but also to offer a novel argument for incompatibilism. Such an argument might run as follows:

(1) PPR: An agent acts freely and responsibly only if she could have refrained from so acting (viz. the dual power to Φ or not-Φ which was exercisable at t).
(2) If determinism is true, then no agent has the powers specified by PPR.
(3) Therefore, if determinism is true, no agent acts freely or responsibly.

There are two ways one might want to resist this argument. One is to dispute (2) by arguing that PPR can be satisfied even if determinism holds. The other is to dispute (1); that is, to show that PPR is false. In what follows, I aim to do just that. More, I will assume that (2) does hold.[4] In so doing, I will defend the coherency of Frankfurt-cases[5] and show that PPR does not provide a strong argument for incompatibilism.

1 Intervention and Control

In this section, I examine the philosophical reasons we might have for thinking PPR is true. My aim is to show that PPR's defenders do not provide a strong argument to convince us of its truth; and that therefore the novel argument for incompatibilism in which PPR is is deployed as a premise is not a strong one.

[4] For discussion of the compatibility of the capacity to do otherwise/refrain with determinism, see Inwagen [27], Lewis [20], Beebee and Mele [6], Beebee [5], and also Inwagen [28].

[5] Matters are complicated here by the fact that Steward, unlike Alvarez, thinks PPR requires powers of token rather than type-refrainment. So, even if it were shown that Frankfurt-type cases are coherent, it may still be argued that, in the actual case, Jones satisfies PPR, and that this is why we hold him morally responsible (see Steward [22] and Wirderker [29]) This response can be dealt with, however, by pointing out that in order to defend the coherence of Frankfurt-cases from this objection one must show that PPR is false. And if PPR is false, then to move from the fact that Jones retains his power of token-refrainment in the actual case to the claim that this is why we say that he is morally responsible will be unmotivated.

Steward and Alvarez appear to offer two arguments. One appeals to the notion of control, the other to the idea that actions are interventions into the course of nature. In this section, I show that whilst both of these ideas certainly isolate key components of our conception of agency, neither supports PPR.

Steward claims that a 'genuine action' requires:

> [...] that an agent possess [...] what I shall call the power of particular refrainment – i.e. the power not to have brought about that very action.

Because:

> It is arguable that actions are, by their very nature, interventions into the course of nature that an agent need not have executed, for that, it might be said, is part of what it is for an event to depend upon an agent's will for its occurrence, and an action is essentially an event that is so dependent.[6]

The argument might be formalised as follows:

(1) Actions are, essentially, events which depend upon an agent's will.
(2) If an event depends upon an agent's will, then it must have been an intervention into the course of nature, that an agent need not have executed.
(3) In order for an event to meet the conditions specified by the consequent of (2), the agent must have possessed the power not to have brought about that event.
(4) (IC) Therefore, AC is true.
(5) (C) Therefore, PPR is true.

There are two problems here. The first concerns the truth of (2), for whilst it is (at least typically) correct that actions are , in some sense, interventions into the course of nature, it is not obvious that this intervention must be such that the agent need not have executed it. On the contrary, there seems to be no intuitive relation between the notion of an action 'depending on one's will' and the notion of not needing to have performed it. For, surely, there are plenty of actions we perform on a daily basis, which are also things that we needed to do. One might say, for example:

(i) I needed to eat, otherwise I would have starved. Or:

[6] Steward [24] p. 84.

(ii) I needed to swim to the surface, otherwise I would have drowned.

However, it seems we can distinguish between two sorts of need, one *hypothetical* and the other *categorical*, where the former sort refer to things one had to do in order to obtain a certain result, and the latter to things one simply could not refrain from doing. This would enable the condition to account for the fact that we do perform many actions that we needed (hypothetical) to perform, whilst still insisting that an event depends on an agents will only if that an agent need not (categorical) have brought it about.

The problem with this is that it is hard to makes sense of a categorical need in any way which does not either (1) make the claim trivially true (and also ineffective), or (2) beg the question against the compatibilist. The first way we might want to make sense of a categorical need is as follows: an agent categorically needed to do something just if there is no possible world in which she could have done anything else. Clearly, however, this makes Steward's claim trivially true: of course events only depend upon an agent's will if there is a possible world in which she could have done other than she did; for it is never logically contradictory for an agent to do otherwise, and hence there is always a possible world in which she does do otherwise. Moreover, if we interpret the notion of a categorical need in this way, then Steward's argument doesn't get her what she wants; for there are possible worlds in which (i) an agent doesn't have the powers specified by PPR, and yet (ii) she doesn't do what she does in the actual world.

The second way to make sense of the notion of a categorical need, then, is to restrict the domain of possible worlds to just those worlds which have same laws of nature as the world the agent is actually in, and which began in the same (micro-)state, and to say that an agent categorically needs to do something just if there is no world within that domain in which the agent does other than she does. This, in fact, does seem to be the way Steward wants us to understand the term 'need' in her paper. What she seems to be asserting is that an event only depends upon an agents will if she wasn't metaphysically determined to do it. The idea, basically, is the familiar one that if determinism is true, then an agent was always going to do that which she does, i.e. it was determined that she would do that which she does, and that therefore she categorically needed to do it. Of course there are logically possible worlds in which the agent doesn't do that which she does in the actual world, but if determinism holds in the actual world, then there is nevertheless an important sense in which she couldn't have done anything else; a sense which is expressed by the claim that she couldn't have done anything other than that which she did in all worlds within the restricted domain.

The problem, however, is that if this is the right way to interpret Steward, then (2) can no longer be an argument for PPR, since, in effect, it says only that for an event to be dependent on an agent's will, and hence to count as an action, that agent must not have been metaphysically determined to do it, (in other words, that determinism must not be the case) – a claim which is basically a statement of incompatibilism, and not an argument for it. Indeed, many philosophers have believed that events can depend on an agent's will even if her bringing about of that event – her performance of the action which quantifies over that event – was causally determined. As Ayer points out, whilst it does follow from determinism that one's actions are the effects of occurrences in the remote past[7], and the laws of nature, etc., it is equally true that they are causes[8]. And to say that an agent causes an event *in the right kind of way*[9] is, plausibly, to say that that action was dependent upon his will. For all that's been said, then, it may be the case that even if one does (categorically) need to bring an event about, one's so doing may still depend upon one's will, and may, therefore, amount to the performance of a genuine action. I now turn to premise (3). (2) states that the bringing about of an event depends upon an agent's will just if his so doing is an intervention into the course of nature. This is correct. But (3) then stipulates that (2) can only be met if an agent possesses the power specified by PPR. I shall now show that this is not obvious.

It might seem that what is meant by 'the course of nature' in this context is the course that nature actually took. But some refection on the matter will show that this is mistaken; for it is actually logically impossible for anybody to change the way the world actually turns out – and this is true whether or not determinism holds. As Dennett puts it:

> It is often said that no one can change the past. This is true enough, but is seldom added that no one can change the future either. If the past is unchangeable, the future is unavoidable, on anyone's account. The future consists, timelessly, of the sequence of events

[7] For an argument to this effect see Van Inwagen [27].

[8] Ayer [4].

[9] If, for example, I am pushed onto a snail and consequently crush it, its demise was not dependent on my will. But if I intentionally crush a snail, then its death would be an event that was so dependant. Of course, the incompatibilist can argue that determinism rules out the possibility of the right sort of causation, but she cannot simply *assert* that this is the case, as Steward seems to be doing here.

> that will happen, whether determined to happen or not, and it makes no more sense to speak of avoiding those events than it does to speak of avoiding the events that have already happened.[10]

What the phrase refers to, then, is the course the world would take from the stand point of a subject with a certain epistemic horizon. That is, it mentions the anticipated future, the way the world will go if we do or don't do such-and-such, or if such-and-such an event does or does not occur. Suppose some astronomers successfully prevent a comet from striking the earth. Undoubtedly, they intervene and change nature's course; more precisely, they avoid the outcome 'a comet hitting earth'. But what they avoid is not the outcome that actually occurred, since no comet did hit the earth – rather, what they avoid is their "projection [. . .] of a certain trajectory into the future"[11]. Now, there is no reason to suppose that an agent in a deterministic universe would not be capable of performing the exact same feats of intervention. Providing they are able to anticipate outcomes, and take appropriate steps to either ensure or prevent their coming into being, then they are able to change the course of nature. Thus, there is no reason to suppose the intuitive idea that an action is an intervention into the course of nature provides any support for PPR.

I now consider the argument from control.[12] In brief, this claims that an agent cannot be said to be in control of his action, in a sense of 'control' necessary for agency, if he lacks the power specified by PPR. What is crucial to this argument, then, is what we are to count as the relevant sense of 'control'. In the context of this debate, it seems that the right sort of control is one which leaves the agent responsive to her environment in such a way that she can be said to count as a genuine agent. This is brought out by the following remark from Steward:

> One might, for instance, insist that part of what is involved in the occurrence of an action is a context in which an agent may be sensibly regarded as having an on-going capacity to do such things as hold up, reverse, or alter the direction and speed of the bodily movement which constitutes the effect of that action, in response to any of a variety of factors [. . .].[13]

[10] Dennett [9] p. 124.

[11] Ibid. p. 125.

[12] See, in particular, Alvarez [1] pp. 77-76.

[13] Steward [25] p. 187

Now, it has recently been argued that the two candidates for the right sort of control are *guidance* and regulative control. The difference can be illustrated by the following example:

Driver: Sally is driving a car which is such that if she attempts to steer right, it will automatically turn left. As it happens, she turns left anyway; the car's mechanism does not intervene.[14]

In this case, we can say that "insofar as Sally actually guides the car in a certain way[. . .], she has 'guidance control', but that insofar as she does not have 'the power to drive her car in a different way', she does not have 'regulative control'".[15]

There is, however, some unclarity attached to this distinction, for, as Steward notes, it is not clear if Sally could have refrained from turning left or not.[16] Since refraining does not require that she turns right, but only that she doesn't (or tries not to) turn left, it's perfectly possible that in the above case, Sally has the power to refrain from turning left.

There are, in fact, three different sorts of control suggested by the example:

Weak Guidance Control (WGC)

To say that Sally had weak guidance control over her 'action' is just to say that she was a (perhaps essential) part of the causal chain which brought that 'action' about. Suppose that for the duration of the action, some evil demon is directly manipulating Sally's brain, in such a way that she becomes merely his puppet, a sort of robot who is utterly incapable of responding to the world, no matter what events occur or might occur. Clearly, she guides the car to the left, and in that sense might be said to be in control – but this is a very limited power indeed.

Strong Guidance Control (SGC)

To say that Sally had strong guidance control over her 'action' is to attribute to Sally *conditional powers* which make it true to say of her that, although, in the event, she does turn left, she nevertheless could have (tried to) not turn left *if* the world had gone a different way. Suppose determinism is true in Sally's world. Then, she was always going to turn left (she was metaphysically

[14] See Fischer and Ravizza [12], Chapter 2.

[15] Ibid.

[16] Steward [23] p. 87.

determined to turn left), and in all worlds which share the same laws of nature and begin in the same (micro-)state as the actual world she still turns left. However, it might still be true of Sally that, if some event, X, had occurred, then she could have responded to it by trying not to turn left. It doesn't matter that X was never going to occur in the actual world, or, for that matter, in all other worlds which share the same laws of nature and begin in the same state, for to attribute to Sally this kind of conditional power is just to mention her *intrinsic properties.* To say of Sally that she has the relevant conditional powers to respond, by trying not to turn left, to a given set of events, F, is just to say if any F-events occur, Sally *would* be able to try not to turn left in response to them. It doesn't matter if no F-events occur, or even if they ever will occur; she is still such as to be capable of responding to them *if they do*. An analogy here might help: suppose that an evil demon grants Tom's wish, and gives him the powers to ride a bike. That's to say, the demon alters Tom's psychological-physiological constitution in such a way that now, unlike before, Tom has the right intrinsic properties which will enable him to be able to ride a bicycle if a bicycle riding opportunity occurs. Being evil, however, the demon proceeds to destroy all the bicycles in the world; and therewith all bicycle riding opportunities for Tom. Clearly, Tom still has the conditional power to ride a bicycle *if the opportunity arises*; and this remains true even though we know that *the opportunity will never arise.*[17] Similarly with Sally: she has strong guidance control over her turning left just because she has the right conditional powers which would enable her to respond to certain events (e.g. a man stepping in front of the car) *if they arose* by trying not to turn left. Since her world is deterministic, she was never going to get the opportunity to exercise those conditional powers, but it doesn't follow from this that she didn't have them.

Regulative Control (RC)

To say that Sally has regulative control is to attribute to Sally a stronger power than the merely conditional power of strong guidance control; it is to attribute to her the power specified by PPR, i.e. the dual power to turn left, or (try to) refrain from turning left which was exercisable at the time of the action. The difference between SGC and RC can therefore be brought out in the following way: The Sally of the SGC example had *conditional* powers which are *exercisable* only in worlds where the relevant events occurred; since

[17] This, I think, is why Lehrer [19] is right to say it is possible that "a man could not have done what he would have done, if he had willed to, chose to, tried to" p. 29. cf. also Moore [21].

the relevant events were determined not to occur in her world, he power was *not exercisable* in her world. The sally of RC, however, has a power the exercisability of which is not conditional on *what the world does* – the Sally of RC has the genuinely exercisable power to refrain from what she does at the time of her *action no matter what*.

Now, the argument from control states that an agent only performs a genuine action if she is responsive to the world in the right sort of way, if she could be 'sensibly regarded as having an on-going capacity to do such things as hold up, reverse, or alter the direction and speed of the bodily movement which constitutes the effect of that action, in response to any of a variety of factors'. Clearly, WGC is not strong enough to satisfy this requirement: to be a cause of an event is not yet to be capable of responding to the world in the right kind of way. It's not obvious, however, that SGC isn't enough. For the agent with SGC has the relevant conditional powers to be such as to respond to Steward's 'variety of factors' if they occur. As we have seen, even if these factors don't occur, she will still have these conditional powers; since to mention these powers is just to mention her intrinsic properties. Unless it is shown, therefore, that the right kind of responsiveness requires RC – something which Steward and Alvarez have not shown – the argument from control does not establish PPR.

Indeed, it seems that SGC really is enough. Suppose that tomorrow, the scientific community announce the discovery that determinism holds. Would they also have to announce that human beings were not, as we previously thought, the kinds of creatures who are highly responsive to their environment? It seems to me that they would not. For our responsiveness is a matter of our being *highly sensitive* to our environment, a sensitivity which is dependent upon our *intrinsic properties*. That's to say, we are the kind of creatures who are psycho-physiologically sophisticated in such a way that if any of a range of possible events occur, we are capable of the appropriate responses. The fact that no events will occur other than the events that are determined to occur does nothing to mitigate the fact that *if* an event of a relevant sort *were to occur*, each of would be capable of responding to it in the appropriate way.

2 Frankfurt-cases

In this section, I examine two arguments which make use of PPR in an attempt to challenge the coherency of Frankfurt's interpretation of Frankfurt-cases. Consider the Frankfurt-case where Black manipulates:

> [...] the minute processes of Jones' brain and nervous system in some direct way, so that causal forces running in and out of his synapses and along the poor man's nerves determine that he chooses to act in the one way and not the other.[18]

One worry Alvarez has about this case is that since 'the 'decision' would not be the outcome of Jones' practical reasoning, Jones does not, in fact, *really decide*, and hence doesn't *really* perform a genuine action. The problem with this, however, is that we have, *prima facie*, no reason to think that Black, in manipulating Jones' brain, couldn't bring about a process of practical reasoning which culminates in a decision. Alvarez's thought, I suppose, runs as follows: An episode of genuine practical reasoning requires choosing between options which are 'genuinely' open. If, however, the outcome of one's reasoning is *determined*, then one's options were not genuinely open, but merely seemed open, 'from the inside'. This mere seeming is simply not enough for a genuine case of decision-making.

The problem is that it's just not obvious that the options one 'chooses' from need to be *objectively* open; perhaps their being *subjectively open*[19] – open 'from the inside' – is sufficient. Let us return to the example of the scientists trying to avoid the outcome 'comet hitting earth'. Suppose one of the scientists reasons like this: "It seems to me I have three options: (a) Go back to bed, (b) Go for one last walk along the beach, or (c) Try to save the planet. Well, (a) and (b) are, on reflection, rather silly ideas, I'd much rather save the planet than go to sleep or walk along the beach, therefore I will choose (c)". Now, if determinism is true, then it was always going to be the case that the scientist would choose (c), his future was not *objectively open*. But it was certainly *subjectively open*; it seemed to him that there were options, and it seemed to him that he chose one of them. Actually, it seems we can say something stronger; it didn't merely seem to him like there were options and that he made a decision, but there *actually were options* and *he really did make a decision*. For perhaps considering an 'option', in the context of practical reasoning, is just to consider an *epistemically possible* (from one's own epistemic horizon) event; and perhaps to decide is just to go ahead with one of these epistemically possible options. For all that's been said, then, Jones

[18] Frankfurt [13] pp. 829-839.

[19] The phrase is due to Dennett [10].

does, in the Frankfurt-case above, make a real decision.

Another worry about this case is that if Black interferes with Jones' brain, Jones is no longer in control of his action in a way which is sufficient for agency. Now, on one reading of this case, this seems right. If, by 'manipulating the minute processes of Jones' brain and nervous system in some direct way', we understand that in the counterfactual case Black would, for the duration of the action, directly manipulate Jones, so as to make him into some sort of puppet, then it may well be that he does not leave Jones in the relevant state to count as an agent. For in determining Jones in this way, he may take away his conditional powers of responsiveness. That's to say, we can imagine Black manipulating Jones in such a way that Jones become oblivious to all sorts of dangers; he walks with a gun in his hand and fails to respond to the presence of policemen; he crosses busy motorways when cars are coming; i.e. he is no longer in possession of *SGC* at the time of his action. For the duration of the action, and the duration of Black's manipulation, it is not true to say of Jones that if an event of a certain sort occurred, he would be capable of responding to it in the relevant way; Black might, for the duration of the action, alter Jones' intrinsic properties in such a way that we cannot say of him that he is a real agent.

That said, it just isn't obvious that Black would need, in the counterfactual case, to intervene in just this manner. Perhaps Black could be more subtle, and merely alter Jones' brain so as to set him on a certain course, a course that would (a) cause him to kill Smith and yet (b) leave him with the relevant conditional powers (intrinsic properties) necessary for responsiveness and agency. For there is nothing inconsistent about the idea that Black manipulates Jones' brain in such a way that he chooses to kill Smith, whilst also making him such that if the right sort of events *interfered*, he would respond to them in the appropriate ways. In manipulating the 'minute processes' of Jones' brain, Black need not render him incapable of the appropriate responses to external stimuli; in fact, he can *determine* him to be so responsive. All he needs to do is ensure that such stimuli do not occur. He might, for example, lock the door of the room they are in, disrupt the mobile phone networks, etc. The mechanics of how exactly Black manipulates the environment need not be spelt out - for we are looking only for a *logically possible* case in which Black determines Jones to act. Thus, so long as it is logically possible – and it surely is – that Black can set Jones' brain on a certain course and then prevent any external factors from interfering, then Black could determine Jones to kill Smith whilst leaving Jones with SGC over his action.

In this counterfactual scenario, Jones would not be able to to refrain from killing Smith – would not have the powers specified by PPR – but would nevertheless retain the relevant conditional powers – the powers of SGC – to allow us to class him as a real agent, and his action as a genuine action.

This discussion brings out an important point. In the literature on Frankfurt-cases, much has been said about what Black has to do to Jones, but very little has been said about what else Black needs to do. But there are other things he needs to do – for if Jones is to have the right sort of control, he needs to be such that if certain factors occur, he will respond in appropriate ways. Thus, if something were to happen which in normal circumstances would cause him to refrain from killing Smith, then he must be such that he would refrain. Now, this does not render Black incapable of determining that Jones kills Smith. All it means is that Black, in so determining, must make sure that no events occur which would cause Jones to respond in such a way that he refrains. Thus, it is not obvious that Frankfurt-cases must be incoherent: on the contrary, they look perfectly coherent, so long as we set up the counterfactual scenario in the right way.

References

[1] Alvarez, M. (2009) 'Actions, Thought-Experiments, and the 'Principle of Alternate Possibilities'', in *Australasian Journal of Philosophy*, 87(1): 61-81.

[2] Aristotle (1984) 'Eudemian Ethics', tr. J. Solomon, in J. Barnes (ed.) *The Complete Works of Aristotle*, Vol. 2. Oxford: Oxford University Press.

[3] Austin, J.L. (1961) 'Ifs and Cans', in J.O. Urmson & G. J. Warnock (eds.) *Philosophical Papers*. Oxford: Clarendon.

[4] Ayer, A.J. (1954) 'Freedom and Necessity', in *Philosophical Essays*. London: Macmillan.

[5] Beebee, H. (2003) 'Local Miracle Compatibilism', in *Noûs*, 37(2): 258-277.

[6] Beebee, H. & Mele, A. (2002) 'Humean Compatibilism', in *Mind*, 111(422): 201-224.

[7] Bobzien, S. (2008) Afterword to *The Philosophy of Aristotle*, in R. Bambrough (ed.) *The Philosophy of Aristotle*. London: Signet/Penguin.

[8] Capes, J.A. (2012) 'Action, Responsibility, and The Ability to Do Otherwise', in *Philosophical Studies*, 158(1): 1-15.

[9] Dennett, D. (1984) *Elbow Room: The Varieties of Free Will Worth Wanting*. Oxford: Oxford University Press.

[10] Dennett, D. (2003) *Freedom Evolves*. London: The Penguin Press.

[11] Fischer, M. (1982) 'Responsibility and Control', in *Journal of Philosophy*, 79(1): 24-40.

[12] Fischer, M. & Ravizza, M. (1998) *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge University Press.

[13] Frankfurt, H. (1969) 'Alternate Possibilities and Moral Responsibility', in *The Journal of Philosophy*, 66(23): 829-839.

[14] Franklin, C.E. (2010) 'The Problem of Enhanced Control', in *Australasian Journal of Philosophy*, 89(4): 687-706.

[15] Honoré, G. (1964) 'Can and Can't', in *Mind*, 73(292): 463-479.

[16] Hume (2007) 'Enquiries: Concerning Human Understanding and Concerning the Principles of Morals', Reprinted from the posthumous edition of 1777. Oxford: Clarendon Press.

[17] Hume, D. (2009) 'A Treatise of Human Nature', in D. Fate Norton & M.J. Norton (eds.) *Hume: A Treatise of Human Nature*. Oxford: Oxford University Press.

[18] Hobbes, T. (1999) 'Treatise: Of Liberty and Necessity', in V. Chappell (ed.) *Hobbes and Bramhall on Liberty and Necessity*. Cambridge: Cambridge University Press.

[19] Lehrer, K. (1968) 'Can's without If's', in *Analysis*, 29(1): 29-32.

[20] Lewis, D. (1981) 'Are We Free To Break The Laws?', in *Theoria*, 47: 113-121.

[21] Moore G.E. (1912) *Ethics*. Oxford: Oxford University Press.

[22] Steward, H. (2006) 'Could Have Done Otherwise: Actions, Sentences and Anaphora', in *Analysis*, 66(290): 95-101.

[23] Steward, H. (2008) 'Moral Responsibility and the Irrelevance of Physics: Fischer's Semi-Compatibilism vs. Anti-Fundamentalism', in *Journal of Ethics*, 12(2): 129-145.

[24] Steward, H. (2009) 'Fairness, Agency and the Flicker of Freedom', in *Noûs*, 43(1): 64-93.

[25] Steward, H. (2012) *A Metaphysics for Freedom*. Oxford: Oxford University Press.

[26] Van Inwagen, P. (1978) 'Ability and Responsibility', in *The Philosophical Review*, 87(2): 201-224.

[27] Van Inwagen, P. (1983) 'An Argument for Incompatibilism', from *An Essay on Free Will*. Oxford: Clarendon Press.

[28] Van Inwagen, P. (2004) 'Freedom to Break to Laws', in *Midwest Studies in Philosophy*, 28(1): 334-350.

[29] Wirderker, P. (1995) 'Libertarianism and Frankfurt's Attack on the Principle of Alternate Possibilities', in *The Philosophical Review*, 104(2): 247-261.

Haecceity As Twofold Negation: A Syncretic Account Of Scholastic Individuation[*]

Peter Damian O'Neil
Heythrop College

Introduction

Individuation concerns itself with the question of what makes this or that particular an individual. This can be exemplified by considering two grains of sand which share the same properties – and asking oneself what the difference between them, if any, is. In this paper I am going to present a syncretic account of a modern view of two opposing theories of individuation, namely those of Henry of Ghent and Duns Scotus. Henry (1217-1293) lived a generation before Scotus (c.1265-1308) and taught that individuation was the product of a twofold negation. This project entails a primary negation of divisibility and a consequent negation of difference – this combination is the "indivision of a thing and its division from all else".[1] I am going to reconcile this project, which was heavily criticised by Scotus in his *Ordinatio* II, with Scotus's own conception of individuation as *haecceity*. *Haecceity* means 'thisness', and is what constitutes being to singularity,[2] it is the "ultimate reality of the being that is matter or that is form or that is composite".[3] *Haecceity* is the operative body between secondary substance (*deuterai ousiai*) and a primary substance (*prote ousia*).[4] In our earlier example of two grains of sand, they would share the same *quiddity* whilst having a different *haecceity* – they are distinct manifestations, 'this & that' of the same 'what'. In this way, the *deuterai ousiai*, or *quiddity*, 'what it is to be a grain of sand' is manifest through *haecceity* into 'this grain of sand', the *prote ousia*.

The central question: can we understand this manifestation, this process of *haecceity* as a twofold *quidditative* and existential negation? There are three main stages in the argument for the affirmative. In the first place, we need to be clear on what we mean by an 'individual', and we will undertake an

[*] Delivered at the BUPS Summer Conference 2012 on 2-3 June 2012 at the University of Leeds.

[1] Wolter [7] p. 23.

[2] Caird [2] p. 196.

[3] Spade [5] p. 107.

[4] The Mediaeval thinkers referred to the *deuterai ousiai* as *quiddity*, which denotes the 'whatness' of a particular.

analysis of the idea of *prote ousia* as 'the individual'. In the second place we will investigate *haecceity* as 'that which makes the individual an individual'. These two stages are principally nominal, and constitute themselves as laying a definitional groundwork in which the third stage will manifest itself. The third stage is the central thesis of the text, combining an understanding of the underlying 'entity' present in the definition of *nomen* by Thomas of Erfurt, with Henry of Ghent's theory of individuation through twofold negation. This third stage is presented in the manner of a reply to the criticism of Henry's theory by Duns Scotus, and draws the earlier stages into itself through negation of Scotus's *modus operandi*, whilst embracing his conception of an intermediate *haecceity* between *prote ousia* and *quiddity*. In this way, we paint a syncretic account of these two thinkers, combining Scotus' 'what it is to be what it is that individuates' with Henry's 'what is it that individuates', to produce an account of individuation.

1 *Prote ousia* as the Individual

The central question for anyone discussing the problem of individuation is, what is an individual? For the purposes of this paper, I shall present an individual as the *prote ousia* which is in real unity. Real unity is one of four main types of unity, pertaining to things which are inseparable from a thing or one another insofar as to separate them disposes of the thing. There are three other types of unity or distinction, the formal, conceptual and intentional. For the purposes of this paper, I shall concentrate solely upon the real, formal and conceptual. A thing is an individual when it has real unity, even if it obtains any of the other three distinctions.

The conceptual distinction is the most simple, it pertains to a distinction existing only in the mind, such as the distinction of the Morning Star and Evening Star. The formal distinction is more complex, Scotus gives an account of this distinction as concerning what he calls *rationes formales*, *formalitates* or *rationes* – which he typifies with the example that:

> x and y are formally distinct or not formally the same, if and only if (a) x and y are or are in what is really one and the same thing (*res*); and (b) if x and y are capable of definition (in the strict Aristotelian sense, in terms of genus and differentia), the definition of x does not include y, and the definition of y does not include x; and (c) if x and y are not capable of definition, then if they were capable of definition, the definition of x would not include y and

> the definition of y would not include x.[5]

In this way, Plato's intellect and will would be formally distinct, whereas Plato's mind would be really distinct from Glaucon's, but only conceptually distinct from Aristocles'. In this way, *prote ousia* exists despite the formal distinction of Plato's will and intellect, or the conceptual distinction of Plato and Aristocles, and a *prote ousia* does not exist between Plato and Glaucon. It is using these distinctions that I hope to show that individuation occurs only through the secondary negation of extension, and though it may pertain to the same *res* as positive entity we can speak of them as distinct in *formalitates* – thus allowing them to immediately reciprocally imply.

2 *Haecceity* as Individuation

It can be shown that *haecceity* is what creates an individual. We shall undertake an examination of the exact meaning of *haecceity*. In this paper, we are drawing upon Scotus's use of the term *haecceity* as 'what it is to be what it is that individuates', rather than his active use of the term as 'what it is that individuates', instead, we are regarding the active *haecceity* as Henry's 'what it is that individuates'. Being careful to note that Henry's theory of individuation came prior to the birth of the term *haecceity*, we are undertaking a syncretic adoption of Henry's 'mode of individuation' into Scotus's later conception of 'what it is to have a mode of individuation'. At this stage, we are examining *haecceity* as the principle of individuation, rather than speaking about the *modus* of *haecceity*. However, this exercise of definition is by no means a simple one, as we have to be exceptionally critical in how we understand *haecceity* in relation to *rationes formales* or accidents.

Therefore we begin our definition by asking what it is that this noun refers to, but primarily – what a noun (*nomen*) refers to. It would be insufficient to define the noun as the signifier of substance, quality or case. Thomas of Erfurt, the most influential of the Modist Grammarians, makes it clear that this account does not work, for the over-simplistic approach to *nomen* would itself exclude negation and fiction.[6] Therefore, we must posit something further, which Thomas does – proposing the property of entity as a solution to this issue. That is not to say however, that *nomen* are without substance,

[5] Kretzmann, Kenny & Pinborg [4] p. 415.

[6] Bursill-Hall [1] p. 155.

quality or case, just that this, too primitive, account would be insufficient. Thomas defines entity as "the property of condition and permanance"[7]. This distinction of Thomas' is critical, because we are understanding *haecceity* beyond substance. If we recall, we are using the term *haecceity* to bridge the gap between *prote ousia* and *quiddity* (*deuterai ousiai*), which are respectively primary and secondary substance. In this way, the *nomen haecceity* functions as a inter-substantive part of speech (*pars orationis*) which exists in relation between the two forms of substance.

Following an understanding of the *nomen* as a *pars orationis* concerning permanance and condition, we can proceed to ask what are the essential modes of *haecceity*, which we shall give in terms of individuation. At this juncture, some logicians might misunderstand and believe that this is simply a presentation of tautology. In this case, it seems that there is a real conjunction obtaining between individuation and *haecceity* – and indeed this is definitional, but it would be an oversimplification to then posit that there was no diversity of *formalitates*. Therefore, I shall demonstrate the essential mode of *haecceity* and how this entails a reciprocal implication and real unity.

Subaltern to the *nomen*, we can make a twofold distinction through the modes of signifying *per modum communis* and *per modum appropriati*; that is to say, through the common and proper modes. It is however, only natural to suppose *haecceity* would only pertain to the proper mode of signifying; which is defined as a property which "is indivisible among several subordinates" – in fact, Thomas himself gives the pronoun "*hic* [he/she/it]" as an example of the proper mode.[8] We can contrast this to the common mode, which pertains to divisibles, this is not of particular relevance to *haecceity* as if we divide something essential to 'this' then it ceases to be 'this' – by all means we may divide accidental qualities, but these are of no relevance to the matter at hand. In this way we can see *haecceity* as the indivisible 'thisness' of particulars, which is what it is to be an individual particular, which is what individuation is – therefore, the *nomen haecceity* really pertains to the individuated object, however, we must not conflate the two, as they are not synonymous – because though *haecceity* is an individuated object, an individuated object is not *per se* a *haecceity*. This is because, we can individuate *rationes formales* of a thing which may pertain to matter, composition or form – which is not what constitutes a *haecceity*, that being the perennial thisness of a thing, upon which

[7] Ibid.

[8] Ibid. p.147.

"we can still distinguish further several formally distinct realities [...]"[9].

Now, the astute observer might believe I am regarding these accidents as mere accidents with regard to the object at hand, and ignoring these accidents *qua res per se* and instead regarding them only insofar as they are accidents. In this way, one might say they have their own *haecceity* insofar as they are called *prote ousia*. This is certainly the case, insofar as we can allow accidents *qua res per se*, which we cannot do in all cases. For example, if we are to posit, an accident, which is individuated from a *prote ousia* by a *rationes formales*, and further posit that this accident cannot exist *qua res per se*, then we are unable to speak of this thing as having a *haecceity*. This is because *haecceity* properly constitutes to singular being, and those things which cannot exist *per se*, and are in real unity and obtaining as a *rationes formales*, cannot be described in terms of a singular being, but merely in terms of an individual *rationes formales*. The *haecceity* therefore, pertains only to to real unities, that is, indivisible bodies, rather than any divisible *rationes formales* of these *prote ousia*.

To clarify this division, I shall note that it is important not to get carried away here, and determine that something divisible has not a *haecceity*, for this would be equivocation of the term divisible. We can proceed by way of example; a tricycle is *per se* divisible into three wheels, but it is not divisible qua *quiddity*, for division of wheels would be a privitation of essential characteristics of what it is to be a tricycle, in the *quidditative* case division would be liquidation. Moreover, division of a prime number is possible, but upon division it would cease to have the *quiddity* of a prime number – so whilst it is divisible as a number, it is not divisible as a prime number. In this sense, atoms are *quidditative* rather than composite or numerical, Scotus himself presents seven arguments for this, the strongest two of which are, firstly, that if every real diversity was numerical, so would every real diversity – and consequently, Plato would be equally distinct from Socrates as from a line, and secondly – all real diversity cannot be numerical, "because in genus there is no singular that is the measure of all that are in that genus"[10]. Therefore, division in the sense used is modally *quidditative*.

[9] Wolter & Frank [8] p. 187.

[10] Spade [5] pp. 60-63.

3 *Haecceity* as a Twofold Negation

Most of what has been said is obvious, the contention arises insofar as we are to ask the question, 'how is it that *haecceity* is?'. This is where Duns Scotus and Henry of Ghent disagree. I shall briefly describe a modern interpretation of Henry's case, and then present in sequence seven of Scotus's arguments, six of whom are from *De principio individuationis*, which I shall draw upon first, and then I shall present the later argument he posits in his *Reportatio*, I shall then discuss each argument in turn. In this way, I hope to answer Scotus's objections, and be able to adopt Henry's theory of the mode of individuation into Scotus's meta-theory of *haecceity* as the inter-substantial medium between *prote ousia* and *deuterai ousiai / quiddity*.

As we know, theories of individuation want to arrive at the individual, the *prote ousia*. Let us start with a safe starting point, the intentionality (*intentio*) of this paper, in order for us to be able to meaningfully interact with this paper it must tend in some way or fashion into the mind of the person receiving it. We must note, this is not a claim about the facticity of the tending, or how well it reflects the world, etc., but merely a claim that in some way, this paper is tending into the mind. How do we speak meaningfully about this paper? Firstly, we have a negation of *quiddity*, we say, this paper is defined by what it is not, we do not say that this paper is the train window, or the hands on the keyboard, in this way – we are making a *quidditative* negation, we are saying – whatever the *quiddity* of this paper, it is certainly not my hands, it does not have those properties. Similarly, geometric figures are defined in terms of negation, a line is as long as it extends until negation, when we negate the line – we can describe how long it is. Then we have the existential negation, we know that this paper is something which is constrained in *quiddity*, but how can we make a claim about this paper being different from say, another exact copy of it. Printing off two copies would not suffice the pettifogger, for they would quibble about particular differences in atomic structure, and so forth – so let us say therefore, that we engineer two exact copies of this paper, in a vacuum. How can we say one is not the same as the other one? We can say paper A is distinct from paper B insofar as it is not paper B, this is the existential negation. If we have a bale of hay, the needle in it is distinct *quidditatively*, from the first negation, whereas each piece of hay is distinct from the other pieces, firstly insofar as they are distinct *quidditatively*, such as one being longer, wider, etc., and secondly, insofar as identical object A is not object B. These two negations constitute individuation, and allow us to speak about the *prote ousia* of individual things. Thus we have the "indivision of a thing and its division from all else"[11].

The first of Scotus's counter-arguments concerns the repugnance of division in individual substance, and how this necessitates positive entity in the individual:

> [...] to be divided into subjective parts is repugnant to an individual material substance. Therefore, it is repugnant to it because of something positive it has, and hence not repugnant because of some negation in it.[12]

However, this is not satisfactory. For to be *an* individual material substance rather than *the* material substance, one must inhere a negation. This is the case because, in every material substance, extension is implicit, and no two extensions of the same mode can be infinite, for this would be a contradiction – therefore, things are not what they are not, as much as they are what they are. Moreover, a line is defined by termination, consisting of two *termini*, through which extension is manifest. On the contrary however, were it the case that we speak of *the* material substance, this can be said to be defined by something positive in it. However, the first negation is division, and division of an integer by itself is wholly one, and the same one that Aristotle speaks of in Book X of the Metaphysics, so to speak of the primary negation of division being incapable of producing singularity of individuation in the instance of a singular extension within a *quiddity* is clearly false. Moreover, a *quiddity* is distinct insofar as it is not extended, this is what individuates entities *in quid* from total entity, furthermore, it is distinct insofar as it is divided *in quid* to the *quiddity* which it is. Regarding this, there is a reciprocal implication in both modes, which are distinct *rationes formales* of the same individuation. A further objection is that negation is only demonstrated *quia* rather than *propter quid*, however this first negation is division, which is a *rationes formales* of *haecceity*, therefore *propter quid* it is certain that the primary negation causes formally and materially *quidditative* individuation, and in the case of motive objects – efficiently.

The second objection that Scotus gives concerns the impossibility of finding formal repugnance through negation, which Scotus holds only to lead to the removal of proximate potency:

[11] Wolter [7] p. 23.

[12] Ibid.

> [...] the cause of any formal repugnance is not to be found in the fact that a negation follows and we have another negation. Though a proximate potency may be removed by a negation (as the proximate potency to see is removed by no objects to be seen, and the proximate potency for division is removed by the lack of quantity), still the reason for a "formal repugnance" has to be something positive.[13]

This is false however, as can be demonstrated. Formal repugnance consists in negation, for what it is to be repugnant to a thing is what it is to be essentially rather than accidentally in a state of negation of form. Efficiently speaking therefore, since repugnance is formal negation, that which is the formal negation of another is itself efficiently incapable of reconciliation, even if the lack of reconciliation is not caused efficiently by other bodies which are formal negations; for this would imply efficient negation, which is not what has been claimed. On the contrary, this negation can be said to be something positive in the bodies; but this something positive is synonymous with the formal negation of the bodies to one another, and is merely asserting a different *rationes formales*. Moreover, the conjunction of the positive with the negative is itself a reciprocal implication, as has been demonstrated here, *quia*.

The third objection Scotus gives is that the hypothetical removal of negation shows that it is necessary to postulate a positive cause of repugnance:

> [...] Hence if the negations were removed per impossibile it would still be necessary to postulate something positive as the cause of the repugnance.[14]

However, that something positive is the essential, rather than accidental nature of the subject of formal negation. On the contrary, this then implies that this mode is itself something positive, a *haecceity* determined positively, rather than through negation. This is true, but is not an objection, as the *haecceity* is a *rationes formales* of entity, *prote ousia*, which we can only speak of as *intentio*, but this is an epistemological question. To postulate something positive, is to look beyond negation, which *intentio* cannot do, as it is finite, anything in *intentio* must be finite, or indeed, if infinite – must be negated to finitude.

[13] Ibid.

[14] Ibid.

The fourth objection given by Scotus is that perfection is the only mode through which imperfection may obtain repugnance, and that as negation is not some perfection, repugnance cannot obtain through negation:

> [...] it is only because of some perfection that what is of imperfection can be repugnant to a thing. But to be divided into many is a matter of imperfection. Therefore, this is not repugnant to an individual material substance except because of some perfection in it. But a negation is not some perfection. Therefore etc.[15]

On the contrary, it may be true that perfection exists as a positive, but it is not individuated by that positive mode. Existence is a perfection, but individuation is negation, and not some perfection, for division is a matter of imperfection. In this way, *in quid* a thing separate from the totality is positive as a thing, but separate as it is not the totality, therefore this primary negation individuates *quidditatively*.

Thus, we can speak confidently of the first negation, being the primary negation of divisibility as being a cause of individuation to have been shown *propter quid*. In this way, we have a *quidditative* individuation, an individuation of whatness's, wherein a *quiddity* is defined by negation of what is extraneous to that *quiddity*. A thing is what it is when what it is is not divided into further things that it could be. The next negation is the division of difference; this is what concerns plurality of quiddities obtaining the same mode. That is to say, what makes two things, which are what they are when not divided into further things that they could be, distinct from one another. In this way what it is to be this *haecceity* and not the other *haecceity* is primarily *quidditative*, but secondarily differential. In the first case, let x be distinct from y, as what it is to be x is not the same as what it is to be y, and what it is to be y is not the same as what it is to be x, and therefore there is a primary negation of *quiddity* and they are individuated by formal division. In the second case however, let x and y share the same *quiddity*, so what it is to be x is the same as what it is to be y; but still we would not say that x is y. Thus follows the second negation of difference, wherein y is not x because to be y is to be not x, and to be x is to be not y. This is not a difference of *quiddity*, but of *haecceity*, these objects are separate not through a division of *rationes formales*, but division of instance – negating other instances of the same form produces singularity of instance within that same form; thus this

[15] Ibid. p. 25.

is the individuating *haecceity* in question.

The fifth objection is that *per se* predication is obtained by what is positive and common to the singular, and that this *per se* predication is not through negation:

> Furthermore, the singular receives the *per se* predication of what is common to a singular; but there is no *per se* predication by reason of a negation [...][16]

But this is not the case, for there is no *per se* division by reason of affirmation, but only though negation. This is demonstrated thus: by affirmation we can only assert a totality, and only individuate through negation. Moreover, asserting a totality is not individuation, for individuation implies division, which is negation. On the contrary, it can be said we do not posit a totality, but individuals through affirmation. This is false, however, as all individuals are *haecceity*, *rationes formales*, *quiddities* and so forth, which are divisions, manifestly determined from the totality solely through means of their twofold negation of divisibility and difference, and not through that something within them that constitutes them, for they are not constitute in the singular, but are constitute in the totality. On the contrary, however. Therefore one need not posit a real unity in the totality, as we see this or that cease, and thus an accident is posited as *quidditative*. But this is false, for these accidents would simply be *rationes formales* of the totality.

Scotus' sixth objection concerns the constitution of entity, and how this does not arise through negation:

> [...] entity is never constituted by a negation, for negation always presupposes something positive in which it is rooted.[17]

This is true but irrelevant, after all, entity is "the property of condition and permanance",[18] which is the totality upon which the twofold negation individuates. Thus, the secondary negation of difference can be spoken of as being the secondary principle of individuation in objects.

The final objection comes from Scotus' later work, the *Reportatio* where he presents the *quidditative* argument that individuals cannot be perfectly known through the idea of species, and that the individual must offer up

[16] Ibid.

[17] Ibid.

[18] Bursill-Hall [1] p. 155.

something positive above the species, or be in non-being:

> If the individual were perfectly known through the idea of a species, then whatever positive that an individual implies is contained in the specific nature. And thus nothing is added over and above a species other than a privation or negation, and thus the individual in itself would be a non-being. And consequently either species and individual would differ in nothing positive; or, if they differed in some way and the individual adds something positive to the *quiddity* of the species, then according to that thing [the individual] would not be perfectly known through the idea of the species, but [would be perfectly known] through an idea proper [to it].[19]

The objection is false however, the individual is perfectly known *in quid* through the negation of species, being the termination of the first negation of division. For, without the negation of species, monism is implicit in the totality of entity. For, something positive is in species, but that something positive is also a *rationes formales* of entity, which is only an aspect through negation of itself – for else it would be formally identical to what it is really and accidentally conjunctive to. Therefore, the being is individual *in hac* only through negation of entity, and *in quid* through negation of species. On the contrary, it is argued that this would cause non-being; but the consequent is false. For it to cause non-being, a negation would be elicited *quidditatively*, which implies being, which is the first and primary transcendental which Scotus agrees precedes all other things,[20] even privation. Consequently, entity is implicit in any negation, unless such a negation negates itself *in quid*, which would be a contradiction, which no one holds. Moreover second, nothing positive is added to a species by individuals, for else by by entering a species, one would not be in the species one entered, and would enter another species – which is a contradiction. On the contrary, this further species is efficiently caused by an individual; however the consequent is false – because the major is inhered, not efficiently caused.

[19] Cross [3] p. 79.

[20] Wolter [6] p. 58.

4 Conclusion

Therefore, when we speak about *prote ousia* we are drawing upon a reconciliation of Henry of Ghent and Duns Scotus' theories of individuation. We accept Scotus' inter-substantial *haecceity* as the 'what it is to be what it is that individuates' with Henry's 'what is it that individuates', to produce a syncretic account of individuation. In this way, if I were to be asked to what makes the rock in front of me different, I would say "ah, the *haecceity*, the 'thisness' of the rock, which is derived by indivision of itself and division from all else – a twofold negation".

In many ways, we have reduced Scotus' side of theory of *haecceity* to the simply nominal element of *haecceity* as the *nomen*, the intersubstantial individuator. However his theory remains indispensable with regards to the way in which it functions as a subtle *nomen* with which we can express Henry's theory of negation regarding real unities, whilst avoiding issues of accidental content. Despite this relegation, Scotus' introduction of the *nomen haecceity* to philosophy, alongside his introduction of the formal distinction, show a clarity worthy of his accolade as the *doctor subtilis*. Moreover, despite accepting Henry's *modus operandi* of individuation, it seems only appropriate to express this in terms of the inter-substantial *haecceity* of Duns Scotus. Therefore, in doing this we can maintain a syncretic account of scholastic individuation, woven into a conception of entity as the totality from which we derive *prote ousia* through utilising the *nomen haecceity* in the mode of a twofold negation of *quidditative* and existential qualification.

References

[1] Bursill-Hall, G. L. (1972) *Grammatica Speculativa Of Thomas Of Erfurt*. London: Longman.

[2] Caird, A. (1948) *The Doctrine of Quiddities and Modes in Francis of Meyronnes*. Ph.D Dissertation (University of Toronto).

[3] Cross, R. (2005) *Duns Scotus on God*. Burlington, VT: Ashgate.

[4] Kretzmann, N., Kenny, A., & Pinborg, J. (1997) *The Cambridge History Of Later Medieval Philosophy*. Cambridge: Cambridge University Press.

[5] Spade, P.V. (1994) 'John Duns Scotus: Six Questions on Individuation from His Ordinatio II. d.3, part 1, qq.1-6', in *Five Texts on the Mediaeval Problem of Universals*. Indianapolis, IA: Hackett.

[6] Wolter, A. (1946) 'Duns Scotus: Opus Oxoniense', in *The Transcendentals And Their Function In The Metaphysics Of Duns Scotus*. St. Bonaventure, NY: Franciscan Institute Publications.

[7] Wolter, A. (2005) *Duns Scotus: Early Oxford Lecture on Individuation*. St. Bonaventure, NY: Franciscan Institute Publications.

[8] Wolter, A. & Frank, W. (1995) *Duns Scotus, Metaphysician*. West Lafayette, IA: Purdue University Press.

Leibniz: Eliminating Cartesian Mind-Body Interactionism and Occasionalism[*]

Dorothy Chen
Columbia University/University of Cambridge

Introduction

Leibniz argues for pre-established harmony by an argument from elimination. He considers Cartesian mind-body interactionism and occasionalism. Rejecting both, Leibniz endorses what he thinks to be the only remaining option: pre-established harmony. This paper will look at Leibniz's arguments against Cartesian interactionism and occasionalism. He refutes Descartes by arguing for the causal isolation of substances, which consists of two claims: (a) each substance has no parts, hence it cannot be affected by a substance other than itself and God; and (b) each substance is causally self-sufficient, and so any inter-substantial causal interaction would be explanatorily superfluous. Leibniz refutes the occasionalists by attacking their conception of God as a performer of perpetual miracles, for it is in contradiction of His divine wisdom to interfere with created substances thus. I will argue that this obscure claim about miracles is really a disguised methodological point about how philosophers should explain phenomena using only notions from the subject discussed, and without recourse to divine wisdom.

1 Against Descartes: Causal Isolation

Leibniz is explicit about the lack of causal interaction between created substances (i.e. all substances excluding God) in his *Primary Truths*:

> Strictly speaking, one can say that no created substance exerts a metaphysical action or influx on any other thing. For, not to mention the fact that one cannot explain how something can pass from one thing into the substance of another, we have already shown that from the notion of each and everything thing follows all of its future states.[1]

[*] Delivered at the BUPS Summer Conference 2012 on 2-3 June 2012 at the University of Leeds.

[1] Leibniz [4] p. 33.

There are two claims here, a negative one and a positive one. The negative claim is that nothing could pass from one substance to another, hence no inter-substantial causal interaction is possible. The positive one is that within each substance is already contained all of its future states, and so even if inter-substantial interaction is possible, it would be explanatorily superfluous since the causal source of all of a substance's past and present properties and future states either come from God or within itself. In the section below I will look at these two claims separately.

1.1 Unity of Substance

Leibniz follows the scholastic tradition in conceiving a substance as a unity containing no parts. The tradition traces back to Aristotle who understood substance as the ultimate subject of predication. In order to test whether or not something is a substance, we just need to (a) be able to attach predicates to it, and (b) be sure that it could not be reduced to further parts (otherwise it wouldn't be "ultimate"). This test makes it impossible for any aggregate to be substance, since any statement involving an aggregate as its subject is analysable in terms of the components of that subject. This Aristotelian framework is what Leibniz had in mind when he insisted on the identity between 'one' and 'being' (Leibniz [5] §2). For Leibniz then, any genuine substance can have no parts. He stays true to this claim even when he later finishes developing his theory of monads. Right at the beginning of *Monadology* he writes: "The monad, of which we will be speaking here, is nothing but a simple substance, which enters into composites; simple, meaning without parts" (Leibniz [8] §1).

Leibniz deduces from this unity of substance the claim that substances are causally isolated:

> There is also no way in which it could make sense for a monad to be altered or changed internally by any other created thing. Because there is nothing to rearrange within a monad, and there is no conceivable internal motion in it which could be excited, directed, increased, or diminished, in the way that it can in a composite, where there is change among the parts. Monads have no windows, through which anything could come in or go out.[2]

[2] Ibid. §7.

Underlying this deduction is an influx model of causality, which Leibniz associates with scholastic philosophers such as Suarez. Under this model of causation, substance A causes a change of state in substance B when a part of A breaks off, is transmitted to B, and attaches itself to B, thereby changing the constitution of B. Something literally flows in from substance A to B, hence the term 'influx'. It follows from this model of causation that if a substance has no parts, then no portion of it could detach in order to affect another substance, nor could there be a gap in the substance as it were to receive external influence, for "monads have no windows". This is why Leibniz could not allow for inter-substantial interaction.

Taken as such, the influx model seems highly implausible. Seeing an angry person may cause me to become angry, without any physical thing breaking off from that person and being transmitted to and attached to me. This view only seems to make sense under some medieval system of physics where for any change to occur, there must be a physical rearrangement of parts. But Leibniz's account of causation seems subtler than this. For instance, he allows for each monad to undergo natural changes. These changes could not be caused by other substances due to Leibniz's commitment to the causal isolation of substances, and so they must be self-causing. I will call causation of this type intra-monadic causation. The worry is that it too is susceptible to the same type of objection Leibniz gives against inter-substantial interaction. This is because each monad is supposed to be simple, and so it cannot have parts to rearrange in order to cause changes within itself.

The key to a more charitable reading of Leibniz's account of causation is a more appropriate understanding of what is meant by 'part' in the influx model. It need not be a literal or physical part. In the case of seeing the angry person and becoming angry myself, we could take the parts to be my mental states. Then we could say that seeing an angry person rearranges my mental states in such a way so as to make prominent the state or part of me that is angry. This could just as well be applied to intra-substantial causation. Each of Leibniz's monads has perceptions and appetitions. Intra-substantial causations occur when a change in its perception results in a change in its appetition or vice versa. Here, we could take each perceptual state or appetition to be a part and say that each change in the appetition of the monad rearranges its perceptual states in such a way that one becomes more prominent than others (as in the anger case above). In other words, the presence of some appetition in the monad rearranges the monad's perceptual states so that it is more aware of one state than all of the others, but as a matter of convention we consider this monad as perceiving just one state, since it is unaware of all of its other perceptual states.[3]

1.2 Truth as Conceptual Containment

To see why Leibniz thinks that the causal source of all of a substance's past, present and future properties are either within itself or from God (and not in any other finite substance), we first need to look at his theory of truth as concept containment. He writes in *Discourse on Metaphysics*, §8:

> Now it is obvious that all true predication has some foundation in the nature of things, and when a proposition is not identical, that is to say when the predicate is not expressly included in the subject, it must be virtually included in it. This is what philosophers call *in-esse*, and they say that the predicate is in the subject. So the subject term must always involve that of the predicate, in such a way that anyone who understood the subject notion perfectly would also see that the predicate belonging to it. This being so, we can say that the nature of an individual substance or of a complete being is to have a notion so complete that it is sufficient to include, and to allow the deduction of, all the predicates of the subject to which the notion is attributed.[4]

Leibniz puts this point more concisely in his *Correspondence with Arnauld*: "[...] the predicate is present in the subject - or I do not know what truth is" (Leibniz [5] pp. 111-112.). For Leibniz, substances are complete concepts. And it is a property of complete concepts that they contain all the properties that has been and could ever be attributed to them. Leibniz makes explicit this point in his Primary Truths, where he says, "the complete or perfect notion of an individual substance contains all of its predicates, past, present, and future" (Leibniz [4] p. 32.). Therefore, epistemologically speaking, we only need to investigate into the nature of one complete concept in order to know everything about it, including everything that will happen to it in the future. Ultimately, Leibniz's claim is going to be more radical than this - he thinks that a thorough understanding of one complete concept will tell us

[3] This view presupposes that a monad possesses all of its perceptual states and appetitions at all times (or in the anger example, a person possesses all of his or her mental states all the time) - its current perceptual state/appetition is just whichever one is active/prominent. One might worry as to whether or not this is too demanding of a requirement for the subconsciousness of the monad.

[4] Leibniz [6] pp. 59-60.

everything there is to know about the world, not just what's in the substance. We will not get into that here.

The metaphysical consequence of this view of truth as concept containment is that the cause for change in the substance must come from within substance itself. The reason we could look at the complete concept of a substance and deduce all of its future states must be because information about those future states and the mechanisms needed to bring them about are already contained in the substance itself. On this point Leibniz writes, "[...] in the soul of Alexander there are for all time remnants of everything that has happened to him, and marks of everything that will happen to him [...]" (Leibniz [6] p. 60.). Leibniz calls this collection of intrinsic features of the substance that is the basis of all its future changes the 'internal principle', and he asserts that it must be interior since an argument, outlined in 1.1 above, shows that no external causes could ever have an influence on the substance (Leibniz [8] p. 269.).

It could be argued that there is nothing in the notion of concept containment that prevents a substance from including in its complete concept a predicate of the form 'is caused to be F by substance x'. In other words, it seems that the claim of truth as concept-containment, when taken on its own, is compatible with causal interactionism. I think this is correct. However, Leibniz could resist the inclusion of predicates of the form 'is caused to be F by substance x' in the concept of a substance because causal interaction by a substance x, that is not coming from itself or God, is for him utterly inadmissible. He makes this point explicit in a section of the text I have already cited in passing, but let me just reproduce it for emphasis:

> It follows from what we have just said that natural changes in a monad come from an *internal principle*, since no external causes could ever have an influence into its interior.[5]

As I have mentioned, the lack of external causes is the result of the argument presented in section 1.1. This still leaves us puzzled as to what Leibniz means by 'internal principle'. I will not attempt to unpack this notion completely, or even adequately, but will try to understand why Leibniz is motivated to endorse it.

[5] Ibid. p. 269.

Leibniz endows each substance with an internal principle in order to guarantee his belief that all substances must be active - “for that which does not act, which has no active force [...] can in no way be a substance” (Leibniz [9] p. 221.). He has three motivations for asserting this. First, each created substance for Leibniz is supposed to mirror God. And since God is active, each finite substance must at least possess a minimal amount of active force too. Second, it has been established that inter-substantial causation is impossible. If, on top of that, substances are not active, they could not bring about changes to themselves, which leaves open only the option that God brings about all the changes a substance undergoes. This could be interpreted as God’s performing perpetual miracles, which is in contradiction with His infinite wisdom.[6] Third, humans must be active in order to be free. Leibniz would like to say that Adam didn’t sin because God decreed that he should; rather, he sinned by his own nature (Leibniz [3] §369.). So unless the substance that is Adam could actively sin by a cause that is within himself, he could not have chosen to sin, therefore could not be held responsible.

2 Against Occasionalism: Perpetual Miracles

In the above section, I have argued for and motivated locating the causal source of change in a substance within itself, and argued for the impossible of causal interaction amongst finite substances. These arguments, I hope, were sufficient to show Leibniz’s rejection of Cartesian mind-body interactionism. In this section, Leibniz’s main argument against occasionalism (i.e. that from perpetual miracles) will be presented as a conversation between Leibniz himself and the occasionalist.

Malebranche’s occasionalism, the type Leibniz is mainly seen as responding to, postulates God as the only causal agent. Mind and body don’t interact between themselves; each mental event is correlated with a physical one because of divine action, and similarly God brings it about that each physical event is in accord with the associated mental event. In other words, finite substances do not have causal powers, they merely provide occasions for divine action.

Though a nice solution to the problem of mind-body dualism presented by Descartes, Leibniz rejects it outright. His initial reason for rejecting occasionalism is one I have already mentioned. In his *New System*, Leibniz writes:

[6] I will have more to say about this point in the next section, when I examine Leibniz’s argument against occasionalism.

> It is quite true that in the strict metaphysical sense, one created substance has no real influence upon another, and that all things, with all their reality, are continually produced by the power of God. But to solve problems it is not enough to make use of a general cause and to introduce what is called a *deus ex machina*. For to do this, without giving any other explanation in terms of the order of secondary causes, is really to have recourse to a miracle.[7]

Leibniz agrees with the occasionalist in ruling out inter-substantial causation. However, he thinks that it contradicts God's infinite wisdom to interfere directly every time an agent performs action, or experiences perception and sensation. Such interference by God would constitute a miracle, but our everyday perceptions certainly don't seem very miraculous.

Bayle responds to Leibniz's above criticism by advancing an alternative occasionalist account:

> [...] it cannot be said that the system of occasional causes, with its reciprocal dependence of body and soul, makes the actions of God into the miraculous interventions of a *deus ex machina*. For since God intervenes between them only according to general laws, in doing so he never acts extraordinarily.[8]

The idea here is that instead of thinking divine intervention as being miraculous, we should just think of it as God's creating and keeping in place psychophysical laws. Instead of visioning a separate act of divine interference every time a mental event is correlated with a physical event or *vice versa*, all that is needed is for God to assign such laws to concurrent classes of physical and mental events. Such assignment needs only to occur once. Thus instead of having a procession of perpetual miracles every time action, perception or sensation occurs, God only needs to perform one extraordinary act.

Leibniz responds by saying that, even in this revised account, divine power is still in play in order to keep such general laws in place, even after the miracle of initial assignment of the laws. He redefines miracles as "something which exceeds the power of created things" (Leibniz [7] p. 205.). Since substances themselves are not causal agents, they could not be causally efficacious in carrying out the causal laws that God has assigned. Hence it would still be

[7] Leibniz [10] pp. 149-50.

[8] Bayle [1] p. 197.

miraculous, in the revised sense, every time a psychophysical law is instantiated, because only God has the causal power to realize them. Therefore, even Bayle's alternative occasionalist account is susceptible to Leibniz's initial critique from perpetual miracles.

Here, the occasionalist could easily reject Leibniz's new stipulation of what 'miracles' mean. Then the occasionalist would not be susceptible to this latter critique. However, I think to do so would be to miss the entire point of the criticism. Leibniz only redefines miracles in order to show that it exceeds the causal powers of a substance to instantiate psychophysical laws. Leibniz writes, immediate after stipulating a new definition of 'miracle':

> It isn't sufficient to say that God has made a general law, for in addition to the decree there has also to be a natural way of carrying it out. It is necessary, that is, that what happens should be explicable in terms of the God-given nature of things.[9]

As Rodriguez-Pereyra points out, this is a methodological point about how philosophy should be done. In explaining worldly phenomena, philosophers should not depend entirely on divine intervention. The underlying theological current in Leibniz's time already dictates that everything that ever happens in the world is brought about by divine wisdom; and so it would be merely repeating a triviality to say that some phenomenon is brought about by divine intervention. The job of philosophers, if they want to say something meaningful, is to figure out how God has constituted the world in a way that is itself causally efficacious. Using the God-as-clockmaker analogy, philosophers when investigating how the clock works should answer using the mechanism within the clock, and not just give the non-explicatory response that "the clockmaker has made it so". I think this is the point Leibniz is making when he claims that "in philosophy we must try to show the way in which things are carried out by the divine wisdom by explaining them in accordance with the notion of the subject we are dealing with" (Leibniz [10] p. 150.). Occasionalism fails this exact methodological requirement and hence should not be accepted as a genuine explanation for mind-body problem.

[9] Leibniz [7] p. 205.

3 Conclusion

If all of the arguments in sections (1) and (2) are sound, Leibniz would have successfully eliminated both mind-body interactionism and occasionalism. This argument from elimination does not get us all the way to pre-establish harmony (unless you think that interactionism, occasionalism, and pre-established harmony are exhaustive of the solutions to the mind-body problem), but it gets us very close to it. As Rodriguez-Pereyra understands it, pre-established harmony consists of three claims:

(a) No finite substance acts upon any other finite substance.

(b) Every non-miraculous state of a finite substance is a causal effect of its inherent active force.

(c) God has set up the mind and the body so that there is a correspondence between their states.

(a) is already established by the argument against Cartesian interactionism in section 1, whereas (c) is pre-supposed in the occasionalist view. Therefore, all that is needed to get from here to pre-established harmony is an argument for (b), which I had scratched the surface of in section 1.2. Its subsequent development I leave for the reader.

References

[1] Bayle, P. (1998) 'Note H to Bayle's *Dictionary* Article, 'Rorarius'', in R.S. Woolhouse & R. Francks (eds.) *Philosophical Texts*. Oxford: Oxford University Press.

[2] Jolley, N. (2005) *Leibniz*. London: Routledge.

[3] Leibniz, G. W. (1952) *Theodicy*. W. Stark (ed.). Indianapolis and Cambridge, MA: Hackett.

[4] Leibniz, G. W. (1989) 'Primary Truths', in R. Ariew & D. Garber *Philosophy Essays*. Indianapolis and Cambridge, MA: Hackett.

[5] Leibniz, G.W. (1998) 'Correspondence with Arnauld', in R.S. Woolhouse & R. Francks (eds.) *Philosophical Texts*. Oxford: Oxford University Press.

[6] — 'Discourse on Metaphysics', in R.S. Woolhouse & R. Francks (eds.) *Philosophical Texts*. Oxford: Oxford University Press.

[7] — 'Explanation of Bayle's Difficulties', in R.S. Woolhouse & R. Francks (eds.) *Philosophical Texts*. Oxford: Oxford University Press.

[8] — 'Monadology', in R.S. Woolhouse & R. Francks (eds.) *Philosophical Texts*. Oxford: Oxford University Press.

[9] — 'Nature Itself', in R.S. Woolhouse & R. Francks (eds.) *Philosophical Texts*. Oxford: Oxford University Press.

[10] — 'New System of the Nature of Substances and their Communication, and of the Union that Exists between the Soul and the Body', in R.S. Woolhouse & R. Francks (eds.) *Philosophical Texts*. Oxford: Oxford University Press.

[11] Rodriguez-Pereyra, G. (2009) 'Leibniz: Mind-Body Causation and Pre-established Harmony', in R. Le Poidevin, P. Simons, A. McGonigal & R. Cameron (eds.) *Routledge Companion to Metaphysics*. London: Routledge.

[12] Woolhouse, R. S. (1993) *Descartes, Spinoza, Leibniz*. London: Routledge.

Is Kantian Constructivism a Coherent and Desirable Doctrine?*

William Mosseri-Marlio
Durham University

Abstract

An attempt to rehabilitate Rawls' commitment to Kant was but one of the aims of the Dewey Lectures. Delivered in 1980, Rawls recast 'justice as fairness' as a Kantian constructivist doctrine, jettisoning the claim that the original position embodied a 'Kantian interpretation' in the process.[1] It was hoped that a deeper understanding of constructivism, a method through which ethical maxims are to be realised, would enable a stronger link between Kant and 'justice as fairness'. However, we shall find those who position themselves as Kantian constructivists either articulate incoherent theories, or leave themselves vulnerable to the same criticisms made of Kant.

My point of departure will be an analysis of the Kantian constructivism set forth by Rawls. Finding fault with the taxonomy Rawls provided regarding what constitutes a Kantian constructivism, we will conclude by arguing the idealisations inherent in 'justice as fairness' are not consistent with the Kantian ethical project. For the justifications such idealisations require will render the constructed ethical principles fundamentally relativist, and therefore in conflict with the universality of the categorical imperative. But attempts to articulate a non-idealised constructivism do not advance the Kantian ethical project insofar as we find them to be merely formal. Thus Kantian constructivism falls afoul the same criticisms of Kant. What is more, we shall find that even these tentative Kantian constructivist theories rest on the same idealisations we found debilitating in Rawls. Constructivism, therefore, does not appear to be a congenial method of ethical enquiry for contemporary Kantian scholars.

*Delivered at the BUPS Summer Conference 2012 on 2-3 June 2012 at the University of Leeds.

[1] Rawls [15] p. 256.

Introduction

Primarily, constructivism provides a wholesale reinterpretation of the notion of objectivity and fact in morality. Rather than accepting the existence of an independent moral order, it asserts that ethical principles are justified through a consent-based reasoning procedure; "apart from the procedure of constructing the principles of justice", Rawls argued, "there are no moral facts."[2] Although, because constructivism holds that ethical principles are only justified when consent between a certain type of agent under specific conditions is reached, moral maxims become tied to, and emanate from, "a suitably constructed social point of view."[3] In the case of 'justice as fairness', this 'mediating position' is represented by the original position.

Rawls tended to contrast constructivism with what he labelled 'rational intuitionism'. Before the emergence of constructivism, rational intuitionism, a position exemplified by Plato and more recently by G.E. Moore, had been the dominant force in moral philosophy. On this view, moral concepts cannot be analysed in terms of non-moral concepts. Ethical maxims are legislated from an independent moral order, and are therefore self-evident reasons for moral motivation.[4] It was Rawls' contention that constructivism provided a sharp contrast to this intuitionist view of morality. Firstly, a thick account of the ethical agent is only present in constructivism - because rational intuitionism posits the existence of an independent moral order that might be perceived, no such account is necessary. Secondly, whereas rational intuitionism suggests that maxims are true or false in virtue of their relation to this independent moral order, constructivism has a more modest account of truth. Ethical judgments, because they are constructed, are not true or false, but 'reasonable' for the agents who are subjected to the constructivism in question.[5]

My exposition of constructivism would be aided through seeing how these comments apply to Rawls. In 'justice as fairness', the contractors within the original position represent a thick account of the ethical agent. The principles of justice apply to those of us in western liberal democracies because the latent ideals within our culture permeate Rawls' account of the original

[2] Rawls [14] pp. 306-307.

[3] Ibid [14] pp. 340.

[4] Ibid. p. 343.

[5] Ibid., Lecture III.

position. Indeed, the principles of justice specifically attempted to provide a solution to an impasse in our recent political culture - the problem of how to find an appropriate rendering of freedom and equality, the work of Locke and Rousseau.[6] Hence the principles of justice apply to us provided we connect with the "history and traditions embedded in our public life" that come to inform the aims of the agents of construction.[7]

Rawls' account of the contractors in the Dewey Lectures is a radical departure from *A Theory of Justice*. Whereas previously the agents of construction were an "outcome of the theory of justice," the causation is reversed in constructivism - for now we start with an account of the agent, and the principles of justice are derived from this conception.[8] The impetus of political philosophy is thus re-calibrated from epistemology to practicality. One must adjudicate how to construct an original position that is appropriate for us, one that reflects our shared conception of the moral agent and the well-ordered society. It is apparent therefore that Rawls' is but one manifestation of constructivism, and variations within this method of ethics can be attributed to discrepancies between accounts of the ideal agent, the mediating position, and the just society.[9] Since his characterisation of these three images in the Dewey Lectures was informed by specifically Kantian notions of freedom, equality, and autonomy, Rawls justified his claim that 'justice as fairness' was a specifically Kantian constructivist doctrine.

1 The Move to Political Constructivism

Before I offer an analysis regarding the feasibility of a Kantian constructivism, I wish to suggest that on Rawls' view, whether a constructivism merited the adjective Kantian was dependent upon the particular characterisation of a constructivism's mediating position.

By the time of *Political Liberalism*, 'justice as fairness' was no longer credited as being a Kantian constructivist project. An inconsistency was observed in the argument set out in the Dewey Lectures between the Kantian elements of Rawls' work, and the new emphasis placed on the pluralism. Since the argument that acting from the principles of justice expressed one's Kantian

[6] Ibid.

[7] Ibid.

[8] Galston [2] p. 496.

[9] For an exposition of another constructivist ethic, see O'Neill [10].

autonomy presupposed the acceptance of Kant's metaphysics, a society composed of citizens with disparate metaphysical views could not countenance such formulation of justice. Kantian constructivism, therefore, accorded autonomy a "regulative role...[that made] it incompatible with political liberalism," and substantive changes made in *Political Liberalism* were made in order to combat this contradiction.[10]

But in spite of the correctives made, Rawls would have to accept that even his political constructivism was a Kantian project. In the Dewey Lectures, Rawls asserts that Kantian constructivism is a doctrine that "specifies a particular conception of the person as *an element in a reasonable procedure of construction*, the outcome of which determines the content of the first principles of justice."[11] Insofar as we see a consistency between the conception of the idealised agent within the original position in Rawls' earlier and later constructivism, we find political constructivism to also be a Kantian project.

But what remains to be seen is whether it was appropriate to label a constructivism 'Kantian' on such grounds. We return to Rawls' schema laid out in the Dewey Lectures. From identifying with his idealised account of the ethical agent, defined as free, moral and equal, Rawls argues that acting upon the principles of justice reflects our nature as an ethical being. Thus the reasonableness of the principles of justice is contingent upon one's association of Rawls' ideal of the ethical principle.

A preliminary criticism of Rawls' taxonomy is that this 'thick' account of the ethical agent is not present in the categorical imperative as a procedure of ethics - we do not need to see *ourselves* in a certain way when considering the merits of a universalised maxim.[12] The categorical imperative rather attempts to invoke a transcendental view of social interaction when we evaluate a moral principle since the legislation of universal law necessitates the view that we are "all ends in systematic connection."[13] We find, therefore, that although Rawls contested that the specification of "a particular conception of the person" is mark of a Kantian constructivist, no such account appears in Kant's work.[14]

[10] Rawls [13] p. 99. This quote originally pertained to Kant's moral constructivism, but it is equally applicable to Kantian constructivism.

[11] Rawls [14] p. 304.

[12] Krasnoff [7] p. 400.

[13] Kant [6] p. 39.

[14] Rawls [14] p. 304.

To understand why there are varying accounts of the ethical agent in Kant and Rawls, we must explore the criticism made by Robert Wolff following the publication of *A Theory of Justice*. The problem Wolff raised was that in striving for universality, 'justice as fairness' abstracted away from everything that is characteristically human. Rawls' theory of justice was therefore not relevant enough to us as finite beings to be legitimate.[15] The response Rawls offered was to illustrate that 'justice as fairness' *was* grounded in human psychology. In the Dewey Lectures, Rawls greatly emphasised that his theory of justice answered a question that had a context: "we take our examination of the Kantian conception of justice as addressed to an impasse in our recent political history."[16]

A retrospective reading of *A Theory of Justice* provides hints that Rawls' aims were somewhat more parochial than Wolff's critique allows. For the project of *A Theory of Justice* was to consider an appropriate conception of justice within the framework of a bounded, or "self-sufficient", polity.[17] This was, on Rawls' account, a departure from Kant's work in an important respect. For in placing the state at the heart of ethics, and therefore making justice the centre of his discussion, Rawls assigned a "certain primacy to the social" that was not apparent in Kant.[18]

2 Kantianism in One Country

The great division between himself and Kant *was* the primacy he accorded to the social, but not in the manner in which Rawls envisaged. Rawls seemed to suggest that 'justice as fairness' differed from Kant's doctrine because the original position is a method through which only considerations of justice, as opposed to any ethical consideration, are to be adjudicated.[19] Still, there is a further manner in which Rawls assigned a greater importance to the social. In responding to Wolff's criticism concerning the *a*cultural and *a*historical nature of the agents of construction, Rawls tied 'justice as fairness' to a particular political unit: western liberal democracies. Unfortunately, the greater emphasis Rawls placed on culture and history in the Dewey lectures led to

[15] Wolff [16] p. 179.

[16] Rawls [14] p. 305.

[17] Rawls [15] p. 4.

[18] Rawls [14] p. 339.

[19] Rawls' position on this issue seems *prima facie* bizarre. For a justification and discussion on this topic, see Hill [5].

criticisms that 'justice as fairness' was "Kantianism in one country."[20]

The inconsistencies within this position are obvious. Although Rawls argued that universality in Kantian ethics is of secondary importance, justifying 'justice as fairness' through cultural practices presented the criticism that Rawlsian constructivism was fundamentally a relativist doctrine.[21] For in grounding the ethical in *our* shared conception of the ethical agent, do we not drift towards the acceptance that moral beliefs are merely contingent upon our cultural idiosyncrasies? This is a position that could not be more diametrically opposed to Kant, but the issues with labelling Rawlsian constructivism as Kantian do not stop here. We might wish to consider how constructivism, through seeking ethical justification through consensus amongst culturally similar people, obscures the role of ethical reflection. Rather than pure practical reason, shared *unexamined* beliefs, regardless of their ethical content, provide the justification for ethical maxims.

But the literature in this area has been rather charitable towards Rawls. John Gray, who coined the term 'Kantianism in one country', suggests that the shift towards grounding the justification for a theory of justice in shared beliefs is not tantamount to "any sort of relativism or pragmatism" of the sort expounded by Richard Rorty.[22] An explanation for this position is given by Samuel Freeman, who argues that "moral constructivism affirms a universal conception of moral objectivity and applies fundamental moral principles to all persons capable of understanding moral requirements *no matter how culturally situated*."[23]

Freeman goes on to argue that because Rawls did not suggest that the meaning of justice is not found in the reflective equilibrium of modern liberal democracies, he is not a cultural relativist: "Rawls thinks [in *A Theory of Justice*] that 'justice as fairness' applies to ascertain the degree of justice or injustice in *any society*, regardless of how people there think of themselves."[24] However, this truly is a bizarre sentiment. For Freeman seems to suggests that reasonable conceptions of justice are contingent upon culture but we can apply *our* standards to others who do not share *our* political history. We therefore see Freeman achieving the remarkable feat of espousing an ethical

[20] Gray [4] p. 3.

[21] Rawls [15] p. 251.

[22] Gray [3] p. 164.

[23] Freeman [1] p. 291 (emphasis added).

[24] Ibid [1] pp. 291-292 (emphasis added).

imperialism from within what is essentially an anti-realist moral framework.

Once we start to antagonise Kantian constructivism with charges of cultural relativism, odd defences such as Freeman's start to emerge. This is of course not surprising when Rawlsian constructivism is simultaneously pulled towards relativism by its culturally dependent nature, but pushed towards universalisability by its Kantian foundations. Still, there remains fruitful avenues to pursue this line of criticism; what if I do not identify with the political culture in which I am situated? Does this mean that the conception of justice that defines my society still applies to me in spite of my dissidence? Why should we prefer a state-centric conception of global politics? Thinking of different ways in which we can divide the world's population might very well render differing conceptions of justice, even within western democratic culture. These comments bear out the flexibility we find in constructivism with respect to the jurisdiction of constructed ethical principles. But this feature of Rawls' method of ethical enquiry conflicts with a fundamental tenet of Kant's moral doctrine - that moral precepts ought to apply universally. A coherent Kantian constructivism, therefore, cannot be defined in terms of its mediating position.

3 Abstractions and Idealisations

I do not wish merely to elicit the characteristics of Rawlsian constructivism that do not cohere with Kant's philosophy. Rather, I want to illustrate how a coherent Kantian constructivist project is impossible without either replaying the empty formalism that some saw in Kant's work or reinstating the questionable metaphysics on which Kant's work rested.[25] At this stage we note how many critics, most notably Kukathas and Pettit, saw Rawls' movement towards a more parochial project, informed by cultural and political history, as Hegelian in flavour.[26] Of course, Rawls did not endorse the conclusions of Hegel's political philosophy, nor did he subscribe to an idealist metaphysic. Nevertheless, in taking the social and cultural landscape as his point of departure, Rawls consented to the Hegelian view that philosophy must be practical.[27]

There is certainly a kernel of truth in this characterisation. However, in

[25] See for example Mill [9] and Wolff [16].

[26] For instance Gray [3], Kukathas & Pettie [8] and Galston [2].

[27] Kukathas & Pettit [8] p. 144.

one fundamental aspect I believe that Kukathas and Pettit's analysis is incorrect. They suggest that "like Hegel, Rawls [was] opposed to the abstract philosophical approach to issues of justice."[28] It might be true to suggest that Rawls concerned himself with building a conception of justice that was applicable to our society. We note how Rawls acknowledged the influence of John Dewey, the American pragmatist, in the opening of 'Kantian Constructivism in Moral Theory'.[29] The question, however, is whether Rawls *succeeded* in creating a practical theory. The answer to this is a resounding no.

The justification of this position does not hinge on the influence Rawls may or may not have exerted on the Democratic party since the publication of *A Theory of Justice*. Indeed, John Gray notes with thinly veiled schadenfreude how trivial 'justice as fairness' has been in American politics.[30] The real issue is that 'justice as fairness' can never be of any practical use at all. Why this is the case lies in the distinction Onora O'Neill draws between abstractions and idealisations. Abstraction, taken strictly, is an innocuous practice. Through bracketing predicates, it is a process of simplification that is aimed at facilitating, for instance, decision making-procedures. Language is a form of abstraction because even "the most detailed describing cannot abolish the indeterminacy of language."[31] All normative principles are likewise abstract and necessarily so. It follows that "abstraction does not make ethical reasoning "either irrelevant or impossible."[32]

Since ethics is "inevitably and properly abstract," it is odd that one of the most deeply considered criticisms of Kantian ethics pertains to the notion that Kant's work is an example of abstract ethical theory *par excellence*.[33] However, O'Neill's comments illustrate this criticism is founded upon a misunderstanding - for it is the idealisations rather than abstractions, that the critics of Kantian ethics attack so vociferously.[34] Bracketing certain predicates is one thing, but it is quite another to either deny certain predicates or introduce predicates that do not reflect reality. When in constructivism we augment certain features of the decision making process that do not reflect everyday life, idealisations result. Of course, it is this aspect of Kantian ethics

[28] Ibid [8] p. 147.

[29] Rawls [14] p. 303.

[30] Gray [4] 'Against the New Liberalism'.

[31] O'Neill [10] p. 208.

[32] O'Neill [11] p. 68.

[33] Ibid. p. 67.

[34] For example Wolff [16].

that critics look upon so disparagingly, and rightfully so. Should Kantians insist upon building an account of ethics that contain unacceptable idealisations, it is difficult to see how Kant's work might be extended into the realm of everyday life.

It is now that we consider whether 'justice as fairness' rests on idealisations. Viewing the well-ordered society as not having interstate relations, positing that there are primary goods that facilitate the two moral powers, and the assertion that agents of construction would be mutually disinterested are all examples of idealisations in Rawls' work.[35] These features of 'justice as fairness' do not simplify - they model a world that we do not inhabit. To therefore draw conclusions regarding the appropriate definition of justice from this model would be inappropriate. As a practical doctrine, therefore, 'justice as fairness' fails.

We also note how some of the least Kantian aspects of Rawls' work lead to these idealisations. Due to the 'veil of ignorance' and un-Kantian characterisation of the agents in the original position as instrumentally rational, an account of the primary goods needed to be introduced. The 'veil of ignorance' is thus punctured in carefully chosen places such that agreement might be reached without bias permeating the principles of justice. Nevertheless, why the 'veil of ignorance' is to be tailored in this manner rather than any other requires explanation. So Rawls argued that these idealisations, of the primary goods and the moral agent, are grounded in *our* political culture. Thus we find that the introduction of an instrumental account of reason necessitates idealisations in Rawlsian constructivism. A more Kantian constructivism might prove more coherent and practical than 'justice as fairness'.[36]

4 Towards a More Kantian Constructivism?

The challenge for Kant scholars, according to Rawls, is to harness the conclusions of Kantian philosophy whilst not relying on either a potentially problematic metaphysical scheme.[37] But this reformulated Kantian constructivism must be substantive enough to generate practical response to political problems we face in reality. Without doubt this is a difficult task, for a coherent Kantian constructivism must not include premises that model a world

[35] Ibid [11] p 72.

[36] O'Neill [10] p. 212.

[37] We note that the problematic nature of Kant's metaphysics generated *A Theory of Justice*. See Rawls [12].

which is not our own. A failure to do so would create idealisations whose justifications are not consistent with Kantian ethics.

Onora O'Neill has sketched an alternative Kantian constructivism that she believes only abstracts. Attempting to avoid idealisations, her constructivism is built with the least determinate elements possible. O'Neill is therefore agnostic about the rationality of the agents in the bargaining procedure, the extent to which they interact, and their moral powers. But the constructivism that O'Neill offers in place of Rawls' project cannot generate practical solutions to problems we experience in everyday political life. Answers to hypothetical questions of the nature Rawls considered in *A Theory of Justice* cannot be given in the context of such indeterminacy - we cannot possibly know what the agents of construction would opt for under these conditions. Still, all is not lost. We can posit the responses to modal questions. Agents would rule out, for instance, "principles of deceit [...] of injury or of coercion. For we cannot coherently assume that all could adopt these principles."[38]

There are two criticisms of O'Neill that I would like to highlight here. Firstly, it is difficult to see how her indeterminate constructivism negotiates the formalism she feared J. S. Mill correctly saw in the categorical imperative.[39] O'Neill's response is that although these principles are somewhat vague, "a range of types of action must be rejected."[40] Although, the truth of this statement is trivial. For who would suggest that a coherent account of Kantian justice could involve coercion, deceit or injury? We have not moved any closer towards a Kantian account of justice that is grounded in empiricism and capable of generating practical principles.

But there is a more serious line of criticism. O'Neill urges us to consider whether "particular Kantian positions in fact rely on unjustified... idealisations."[41] She prompts us to evaluate only particular positions because an interpretation of Kant that is given suggests no such idealisations exist in the doctrine of the right.[42] However, we have reasons to offer a transcendental argument regarding the persistence of idealisations in constructivist ethics. If O'Neill's definition of an idealisation is the "[privileging] of certain sorts of human agents and life," on this definition, any manifestation of

[38] O'Neill [11] p. 78.

[39] Mill [9] p. 254.

[40] O'Neill [11] p. 78.

[41] Ibid. p. 72.

[42] Ibid. pp. 74-78.

constructivism will rest on idealisations.[43] Taking O'Neill's constructivism, in basing our theory of ethics on universalised principles, do we not privilege cosmopolitan ethics? In necessitating agreement amongst a plurality of agents, do we not privilege democratic political theorising? In valuing practical reason, do we not privilege our own species?

If my first criticism of O'Neill stands, then because of the problems of idealisations in ethics, a non-formal and coherent Kantian constructivist is impossible without the reintroduction of Kantian metaphysics. We recall that Rawls made the suggestion "to develop a *viable* Kantian conception of justice the force and content of Kant's doctrine must be detached from its background in transcendental idealism."[44] My treatment of constructivism has render Rawls' desired project impossible. Still, if the second criticism stands, constructivism as a method of ethics is no longer available to Kantian scholars. If all constructivism rest on idealisations, and these idealisations require political justification, then universal maxims could never be coherently justified through constructivism. Constructivism, therefore, is not the meta-ethical framework through which the Kantian ethical project can be realised.

References

[1] Freeman, S. (2007) *Rawls*. London: Routledge.

[2] Galston, W. (1982) 'Moral Personality and Liberal Theory: John Rawls' "Dewey Lectures"', in *Political Theory*, 10(4): 492-519.

[3] Gray, J. (1990) *Liberalisms: Essays in Political Theory*. London: Routledge.

[4] Gray, J. (1995) *Enlightenments's Wake: Politics and Culture at the Close of the Modern Age*. London: Routledge.

[5] Hill Jr., T. E. (1989) 'Kantian Constructivism in Ethics', in *Ethics*, 99(4): 752-770.

[6] Kant, I. (1993) *Foundations for the Metaphysics of Morals*. J. W. Ellington (trans.). Cambridge: Hackett Publishing Company.

[43] O'Neill [10] p. 210.

[44] Rawls [12] p. 265.

[7] Krasnoff, L. (1999) 'How Kantian is Constructivism?', in *Kant-Studien*, 90(4): 385-409.

[8] Kukathas, C. & Pettit, P. (1992) *Rawls: A Theory of Justice and its Critics*. Cambridge: Polity Press.

[9] Mill, J. S. (1974) *Utilitarianism and Other Writings*. M. Warnock (ed.). NY: Meridian.

[10] O'Neill, O. (1989) 'Constructivism in Ethics', in *Constructions of Reason*. Cambridge: Cambridge University Press.

[11] O'Neill, O. (2000) *Bounds of Justice*. Cambridge: Cambridge University Press.

[12] Rawls, J. (1977) 'The Basic Structure as Subject', in *American Philosophical Quarterly*, 14(2): 159-165.

[13] Rawls, J. (1993) *Political Liberalism*. NY: Columbia University Press.

[14] Rawls, J. (2001) 'Kantian Constructivism in Moral Theory', in S. Freeman (ed.) *Collected Papers*. London: Harvard University Press.

[15] Rawls, J. (2010) *A Theory of Justice*. New Delhi: Universal Law Publishing Company.

[16] Wolff, R. (1990) *Understanding Rawls*. NJ: Princeton University Press.

Intuitions About Harm and the 'Experience Requirement'*

Thomas Quinn
University of Sheffield

Introduction

> "An individual can be harmed by something only if he has an unpleasant experience as a result of it [...]."[1]

Fischer calls this principle the 'Experience Requirement' (ER).[2] It appears to be widely accepted that this requirement is questionable. Nagel claims that ER leads to the implausible consequence that "even if a man is betrayed by his friends [. . .] none of it can be counted as a misfortune for him so long as he does not suffer as a result"[3]. For Nagel, this consequence of ER is a *drastic restriction*.[4] To him, it seems intuitively wrong that a person isn't harmed by a betrayal, just because she doesn't have an unpleasant experience as a result of it. Other philosophers share this view – for example, McMahan claims that these kinds of cases "constitute counterexamples to [ER]"[5].

It does seem difficult on an intuitive level to bluntly deny that these situations are harmful to the person involved. When considering an example such as Nagel's, it seems wrong to claim that the person's life is going well, even if she is unknowingly ridiculed by her friends. Do these thoughts show that ER is false? I think not. I will argue that appealing to our intuitions in order to refute ER is unhelpful. These examples appear to refute ER but the way that we form our intuitions and judgements about them is flawed. In reality, they tell us nothing about the truth or falsity of ER. In other words, they tell us nothing about whether or not these situations can be harmful for the person involved.

*Delivered at the BUPS Summer Conference 2012 on 2-3 June 2012 at the University of Leeds.

[1] Fischer [1] p. 342.

[2] Ibid.

[3] Nagel [4] p. 4.

[4] Ibid.

[5] McMahan [3] p. 33.

1 Direct and Indirect Unpleasant Experiences

An important first step is to examine the different types of unpleasant experiences that a situation can result in. I am going to distinguish between two different types of unpleasant experience.[6] The 'direct' type of unpleasant experience is the one which is typically due to the person's discovery of the situation. If Pat discovers that she has been betrayed by her friends, she will be upset or experience some negative emotions. The unpleasant experiences induced by the prospect of a situation also count as direct. If Pat somehow discovers that she will be betrayed at some time in the future, she will experience the same kinds of negative emotions. 'Indirect' unpleasant experiences are those experiences that a person has as a result of the other consequences of the situation. Pat's friends betray her by spreading rumours about her. This results in her losing her job, which is an unpleasant experience for her. An important difference between the two is that direct unpleasant experiences can result from a situation only if the person is aware of the situation, whilst indirect unpleasant experiences can result whether or not the person is aware of the situation.

These two types of experience, which may be results of an event, are completely independent of each other. An event may be indirectly unpleasant for a person but not directly unpleasant. In the example of Pat's friends spreading rumours, this would be indirectly unpleasant for her whether or not she ever found out that she lost her job because her friends had betrayed her. Conversely, an event may be directly unpleasant but not indirectly unpleasant. Pat tells Tim a secret. Tim tries to send Jane a text message about the secret, but accidentally sends the message to Pat instead. Pat is upset upon discovering Tim has betrayed her in this way – even though there are no other unpleasant consequences of Tim's betrayal. It may also be the case that both types of unpleasant experience result from the same event.

There are three main points that I wish to draw from this analysis:

(1) In cases such as the betrayal example, which are given as counterexamples to ER, the direct unpleasantness can occur whether or not there are any other consequences of the event. Further, the direct unpleasantness can occur whether or not the event actually takes place at all, since even the

[6] This distinction is very similar to one used by Fischer ([1] p. 342) Whilst he makes a distinction between two ways in which a situation can result in unpleasant experience, I make a distinction between two different *types* of unpleasant experience. The reason for this is that is makes my discussion of examples clearer.

prospect of the event is enough to give a person unpleasant experiences.

(2) Because of these properties, in many cases the direct unpleasantness is not a result of the betrayal itself. Since the direct unpleasantness can occur whether or not the betrayal actually occurs, the unpleasantness is not a result of the betrayal, but a result of certain beliefs – a person's belief that they are going to be betrayed, or that they have been betrayed. This is the case whether or not they are true beliefs, or even justified beliefs.

(3) I can have direct unpleasant experiences without even having the belief that the betrayal will be unpleasant for me in any other way – this is illustrated by the example where Pat is betrayed by Tim.

2 Why Beliefs about Betrayal Lead to Direct Unpleasant Experiences

In the next section, I will explain how these points are important to my argument. Firstly, I will discuss a question that is raised by the aforementioned claims: why does the belief that I have been betrayed make me unhappy, even in cases where I do not also have the belief that I will have an unpleasant experience as a result of it?

Having the belief that I have been betrayed behind my back leads to certain other beliefs. For example, Pat's belief that Tim has attempted to share her secret leads to the belief that Tim is capable of betraying her, that she cannot rely on Tim or trust him, that he has some reason not to respect her wishes and so on. Even though in some situations these beliefs may not be negative, a belief such as this will always be accompanied by a set of beliefs about the general state of affairs surrounding the situation.

In almost every case where a person is betrayed, either the betrayal itself or the state of affairs which surrounds the betrayal will result in an unpleasant experience for that person. It is easy to imagine how the situations I have mentioned are likely to result in unpleasant experiences for the person. For example, if Pat cannot trust Tim then she can never rely on him helping her, which means that she will have to struggle alone in situations where she requires help. So even in cases where the betrayal itself does not result in an unpleasant experience, it is very likely that some aspect of the state of affairs which surrounds the betrayal will result in unpleasant experiences for the person who is betrayed.

I think that it is so likely that there will be some unpleasant experience for the person who is betrayed, and it is so often the case that there are these unpleasant experiences when a person is betrayed, that an association is formed between the betrayal itself and unpleasant experience. We have the emotional reaction, such as fear or sadness, upon finding out that we will have an unpleasant experience, because we are naturally averse to unpleasant experiences. By their nature they are unpleasant, and if something is unpleasant, we wish to avoid it. We have the emotional reaction (the direct unpleasant experience) when we believe we are going to be betrayed because we associate betrayal with unpleasant experience.

This association is not a conscious belief – 'If I have been betrayed I will have unpleasant experiences' – but is on a much deeper level. The association between betrayal and unpleasant experience seems unconnected to the actual facts about the circumstances of a case. It is an entirely general association due to the regularity with which some aspect of the circumstances surrounding a betrayal will lead to unpleasant experience.[7]

Even though it is to some extent involuntary, this disposition to make this association appears to be a useful one in the sense that it would not be best for us to try and eliminate it. If I am concerned with avoiding indirect unpleasant experiences, it is better to have this entirely general and unchanging reaction to events such as my betrayal rather than to try and evaluate the situations on a case-by-case basis. The direct unpleasant experiences, which are a result of my belief that I will be betrayed, mean that whenever I consider what being betrayed will be like, my thoughts are negative. This leads me to avoid betrayal altogether, which is surely a better way to avoid indirect unpleasant experiences than occasionally 'risking it', and not attempting to avoid cases of betrayal, where I think there is a possibility I won't have any indirect unpleasant experiences.[8]

To show that this applies to other cases, consider another example. A famous artist, who has become stranded on a desert island, receives a message

[7] It has been pointed out to me that this relates to Hume's views on the workings of the mind. In the *Treatise of Human Nature*, Hume argues that "there is a secret tie or union among particular ideas, which causes the mind to conjoin them more frequently together, and makes the one, upon its appearance, introduce the other" (Hume [2] p. 662). Of particular relevance is his view that "there is no relation, which produces a stronger connexion in the fancy, and makes one idea more readily recall another, than the relation of cause and effect betwixt their objects" (Ibid. p. 11). A similar sort of idea can be seen in my argument that since a betrayal usually results in an unpleasant experience for the subject; this leads to the subject making an association between betrayal and unpleasant experience.

[8] Perhaps there could be some sort of evolutionary explanation of why we have this disposition. It may be the case that those who have this disposition to associate events with their general outcomes rather than evaluating them on a case-by-case basis survive better.

telling her that since she has been missing, everyone has discovered that all of her paintings are just copies of the works of another, unknown, artist. She knows that this morning she accidentally ate a poisonous fish and that she is sure to die before any ships can make it to the island to rescue her. Still, this news makes her unhappy – she has direct unpleasant experiences as a result of the belief that her character has been denounced. This is due to the association between being exposed as a fraud, and having unpleasant experiences. The event of one's being publicly exposed as a fraud entails certain other states of affairs – one's friends and family know that she has lied to them, one's integrity being brought into question, being the subject of general mistrust. In a normal situation, it is extremely likely that there will be unpleasant experiences as a result of this general state of affairs necessary for the denouncement to take place. This is the reason for the association. However, the association has become so strong that merely the thought of being exposed as a fraud itself results in sadness, apprehension or some other unpleasant emotional state. This is why the artist has the direct unpleasant experiences, even though she knows she will not suffer any indirect unpleasant experiences as a result of the event.

3 The Problem with Counterexamples to ER

Now to consider how these ideas apply to the way we consider the 'counterexamples' to ER.

A number of counterexamples rely on the fact that when faced with a situation where they are going to be betrayed, but will not have any unpleasant experiences, people would still prefer not to be betrayed. Using the ideas I have previously mentioned, this preference can be explained. The person has the belief that she will be betrayed and, as I have shown, the belief that I will be betrayed is sufficient for me to experience direct unpleasant experiences. I think that the fact that the person has the direct unpleasant experiences is what explains the person's feeling that the betrayal will be harmful for her – which in turn explains her preference.

The problem is that the person cannot consider the prospect of the betrayal without having this unpleasant experience. As I have previously mentioned, the association between betrayal and indirect unpleasant experience is deeply ingrained and difficult to get rid of. Since this is the case, she cannot adequately consider what her feelings will be in the future at the time when the loss occurs. Her judgement is affected by the unpleasant feelings she has now.

She cannot separate the thought that she will suffer a loss from the unpleasant emotions which result from it, and so decides that the loss will be harmful to her.

Our intuitions about Nagel's case can be explained in much the same way, even though there is one important difference – when I consider Nagel's example I do not have the belief that I will be betrayed. This may initially seem to be too much of a leap. I have shown that my having the belief that I will be betrayed is sufficient for direct unpleasant experience. However, does this extend to my beliefs about hypothetical cases? Examining the beliefs and processes involved in evaluating the examples will show that the same direct unpleasant experiences result in this case as well.

The important thing to remember is that the direct unpleasant experience is due to the association between betrayal and unpleasant experiences. The belief itself is sufficient but not necessary for direct unpleasant experiences. When I have the belief that I will actually be betrayed, I have the unpleasant experiences due to the fact that the contents of the belief involve the thought of my being betrayed – and the association between betrayal and unpleasant experiences. This happens even in cases where I also believe that I will not have any indirect unpleasant experiences as a result of the betrayal. So it does not matter that in the case where I am evaluating the fictional example, I do not believe that I will actually be betrayed. I do not need this belief, or the belief that there will actually be unpleasant experiences for me. Rather, all that is needed is the thought of me being betrayed. And in order to consider the example, I must entertain the thought of me being betrayed.

This prevents us from evaluating the cases in question properly, and explains why we cannot trust the intuitions that we have about these cases. Because we have these direct unpleasant experiences, we associate betrayal with harm in a way that is not based on rational consideration of the case.

This also applies to the way in which we evaluate the situations of other people, and evaluate examples where another person is the one who is the subject of these events. I feel sympathy for someone who is being ridiculed behind her back because the way that we sympathise with people is by imagining ourselves in their place. This leads to me considering what it would be like if I was being ridiculed behind my back, and this leads to the same factors I have already mentioned influencing my judgement of the case.

If my argument in this paper is indeed correct, the widely accepted 'counterexamples' to ER, such as Nagel's betrayal case, do not successfully refute

ER. Even though it intuitively seems that events such as betrayal are harmful even when there are no unpleasant experiences as a result, and we would prefer not to be betrayed even if we know there will be no unpleasant experiences as a result, these judgments stem from our unconscious association of betrayal and unpleasant experiences, and the directly unpleasant emotions we have when considering these cases. I am not claiming to have shown that ER is correct, merely that appealing to intuitive arguments about the principle is not useful since our intuitions about these cases are not reliable.

References

[1] Fischer, J. M. (1997) 'Death, Badness, and the Impossibility of Experience', in *The Journal of Ethics*, 1(4): 341-353.

[2] Hume, D. (1975) *A Treatise of Human Nature*, (2nd ed.) L.A. Selby-Bigge (ed.) & P.H. Nidditch (rev.). Oxford: Clarendon Press.

[3] McMahan, J. (1988) 'Death and the Value of Life', in *Ethics*, 99(1): 32-61.

[4] Nagel, T. (1979) *Mortal Questions*. Cambridge: Cambridge University Press.

What is the Relationship between the Priest and the Ascetic Ideal?

Robert King
University of Leeds

Introduction

On the Genealogy of Morals (GM) essay III stages a discussion on the relationship between different types of man and the ascetic ideal (AI). One component of this discussion is the relationship between this ideal and the priest. As one consistently discovers with Nietzsche, simply reading the text segregated from the rest of the work is not sufficient to provide us with a full answer to our question. Therefore in order to furnish an answer to the eponymous question of this essay special attention will be paid to GMI & III along with the work of Owen [2] and Ridley [3].

The point of contention in Nietzsche's tale to be addressed in this essay is the appearance of a possible contradiction between three statements: the priest (i) is a creature of ressentiment, (ii) needs the ascetic ideal in order to live, and (iii) uses the ascetic ideal to gain power. The problem is the following: if (i) and (ii) (as one may initially assume) together equate to the priest believing his teaching of AI, then he would see no value in worldly power, leaving Nietzsche without a motivation on which to mobilise the priest to (iii). This essay attempts to show that no such contradiction exists. Rather, what causes the problem as first perceived is ignoring the priest's nobility and mistakenly positing him as like the slave in his relation to ressentiment.

This essay will continue as follows: Firstly will be a definition of the two forms AI takes, followed by an explication of the reading which leads one to decry Nietzsche's account as contradictory. There then follows a subsequent investigation into the priest's nobility and his role in the second phase of the 'slave revolt', and finally a clarification of why no contradiction exists.

1 The Ascetic Ideal

In GMIII Nietzsche concerns himself with the following question: 'what is the meaning of ascetic ideals?'. In so doing he develops two strands to AI;

these we shall call ascetic procedures and AI proper. Learning to differentiate between the two of these is key to finding a resolution to the contradiction.

Ascetic procedures, Nietzsche tells us, are used in order to affirm life (GMIII, §7). The philosopher, for example, may consider himself best able to achieve his desired goal by following a path which has him reject certain other desires (§1, §7) – he may choose to abstain from sexual intercourse, or commit himself to a life of solitude *et cetera*. The key consideration is that following ascetic procedures need not entail any kind of belief in a 'metaphysics of asceticism' - it does not require belief that such a lifestyle is divinely ordained. Moreover, this was the original form of AI, for it is the use of ascetic procedures that Nietzsche talks of in GMI §6 in his discussion of 'priestly aristocracies' who deny themselves 'certain types of foods which cause skin complaints. . . '.

AI proper, on the other hand, involves the above ascetic procedures, and also entails just the kind of belief the above does not. AI proper is AI employed as a meaning for suffering: firstly an omnipotent God who is the font of all values is posited, thus explaining suffering as born of an inability to live in accordance with these values. Secondly, a chance of reprieve from suffering is granted by supposing that this life 'functions as a bridge' (§11) to the next life. It thus allows for a redirection of the feeling of ressentiment (the feeling of frustration which comes about when one experiences a lack of ability to manifest one's will) from an outward feeling (one directed towards those who prevent one from expressing one's will and are the cause of one's suffering), to an inward one. Through AI proper, suffering becomes seen as one's own fault, and consequently all blame rests with oneself. The 'meaning' for suffering it provides is why AI proper is taken up by those who Nietzsche informs us are desperate to find a meaning for their suffering (§15), the slave class.

2 The Contradiction

We must now set out Nietzsche's characterisation of the priest's engagement with AI which might lead one to believe the contradiction exists. The characterisation can be split into three main parts and doing so will facilitate easier assessment of exactly what such a characterisation entails:

(i) The priest is clothed as a creature of *ressentiment* (§15 talks of the priest as being, himself, 'sick'). As Nietzsche places much emphasis throughout GM on the slave being the sufferer of *ressentiment* and this being one of the characteristics which allows one to differentiate between the slave and the noble (the noble being one who can make his will felt, whilst conversely, the slave forever finds his will confronting insurmountable obstacles – namely the noble[1]) it would be reasonable to draw the conclusion here that Nietzsche wishes to identify the priestly class with the slave.

If we proceed on this assumption, combining it with another of Nietzsche's pronouncement (this one engendering the first way in which he sees the priest employing AI), (ii) the priest needs AI to live (§11; "His right to exist stands and falls with this ideal"), it would be reasonable to conclude that the priest requires AI in the same way that the slave does. That is, he requires AI proper as a means of redirecting his *ressentiment* inwards, of positing himself (due to his transgressions) as the cause for his suffering. If this were the case, then it would be necessary for the priest to *believe* in AI proper; it would require a belief that this life is truly a 'bridge' and that all worldly endeavour is merely a prequel and preparation for that which is to come. Moreover, it would require the priest to believe AI proper in the same way as does the slave.

The contradiction makes an appearance when a final statement about the priest is considered; (iii) The priest uses AI to gain power as part of his on-going war against the knightly nobles (GMI); he uses AI to master 'life itself' (§11). Stated more fully, the priest uses AI as a means of inaugurating himself as 'shepherd of the 'herd''. He is able to gain power over the slaves through acting as the conduit through which they may have access to God; the God he creates to provide for the slave that which he has long desired – a meaning for his suffering.

[1] This might, at first, sound a slightly contentious statement. Some might like to highlight that Nietzsche appears to assign the cause of the slaves suffering to a deficient physiology. I would argue that this is certainly *a part* of what causes the slave to suffer and what results in his *ressentiment*; however, it is not the whole story.

The slave would only suffer from his weakness if there were something that his weakness (his deficient physiology) disallowed him. Nietzsche has it that the real problem with being a slave is that his weakness prevents him from expressing himself, it prevents him from acting. Now, how might there be weakness, if there is not an obstacle which is failed to be overcome? How might there be a slave if there is no 'master'? Nietzsche's very language surely directs us to understanding that this obstacle and this master is what the noble is in the early stages of the history Nietzsche plots. Consider, for example, Nietzsche's repeated use of the adjectives 'downtrodden' and 'oppressed' throughout GMI to describe the slave caste. The noble might be this obstacle and this master because he is so overflowing with strength and so able to impose himself on the world as one who coins values.

It is in the slave's nature to be used because of his weakness. On the one hand, the slave is used by the nobles; they constitute the caste that Nietzsche re-casts as the 'birds of prey', who feed on the slave 'lambs' (§13). On the other hand, the slave is used by those that Nietzsche calls the 'opponent', whom the 'powerless will attack' (§10). The slave is even enslaved by the priest, as I will show in this essay.

Here the point of contention manifests itself; if from (i) and (ii) we take it to be that the priest believes in AI then it would be contradictory for him to act according to (iii). For in believing AI he would not place value on this world, rather exactly the opposite, he would renounce all worldly things, for that is what AI proper requires - and so the priest surely would not care to gain power in this world. That is to say that if our conclusion from (i) and (ii) holds, then there is a motivational deficit for the enactment of (iii).

3 Nobility

To escape this contradiction it is necessary to consider the way in which Nietzsche introduces the priestly class in GMI. As Owen [2] points out he is fundamentally a noble with bad conscience; what does this mean?

To be noble is to be the one who 'coin[s] the names of values' (GMI, §2), one whose values originate with a recognition of oneself as 'good'. It is also, as Ridley [3] states, to overcome death in the 'noble way'. Whereas the slave sees death as an end, the noble is able to overcome death because, due to feeling his mode of valuation self-justified, he sees that in his life he has given the poets something to sing about (§2), and so his legacy (if not his person) survives his death. The slave, on the other hand, lives a life which forever comes up against resistance and so has no recourse to overcoming death in this way. As well as these similarities that the priestly noble shares with his counterpart in nobility, the knightly noble, the priest differs in one key feature. Unlike the knightly noble, the priest is not an overtly physical being (§7). This forces him to adopt a different reaction to repression from the knightly noble. He is denied the 'macho' reaction, the overcoming through a brute display of physicality; rather he is required to develop an 'inwardness' (Ibid.). The denial of a physical expression of his will and how it results in the development of an inwardness and intelligence is clearly on display below:

> Priests are, as is well-known, the *most evil enemies* – but why? Because they are the most powerless [Here we might replace powerless as weakly physical when we consider this passage in the context of the rest of Nietzsche talk of the priest and that with which he is compared – the noble] From powerlessness their hatred grows to take on a monstrous and sinister shape, the most cerebral and most poisonous form. (GMI §7.)

It is through this development that he becomes a creature of ressentiment and bad conscience and through this that he might take his '*most intelligent revenge*'. As Nietzsche points out (GMIII, §10) the priest was required to adopt ascetic procedures in order to escape the contempt of the knightly nobles; he had to disguise himself as the 'contemplative man', adopting ascetic modes of life. Though as he is still of the noble caste (and so believes himself to be a rightful source of values), this *ressentiment* does not build up in the same way that it does for the slave and so he does not require AI to overcome death. He needn't believe AI, for he still views himself primarily as 'good' (§6). Rather than using AI proper to give meaning to his suffering as the slave must to re-direct his *ressentiment*, the priest is able to use AI to express his will, i.e. to be victorious in his struggle with the noble (as we will see).

His development of inner resources means that the priest is able to exploit these resources, in affirming his will in leading the second phase of the slave revolt. It also means that he gains power over the slaves and simultaneously dissolves the knightly caste (his long-term rival)[2].

4 Slave Revolt

The second phase of the slave revolt is the move which sees *ressentiment* turned inwards through AI proper. As one whom, himself, suffers *ressentiment* the priest both sees it for the dangerous feeling it can be (Nietzsche states that it can express itself in dangerous explosions – GMIII, §15) and understands what is needed to combat it. The way that he sees of resolving the problem of *ressentiment* (though Nietzsche tells us in GMIII that it is a neither satisfactory nor desirable solution) is by developing his ascetic procedures through incorporating an 'otherworldly element' (Ridley [3]). In creating the otherworldly aspect of AI, he changes it from merely procedural to doctrinal (he creates AI proper), and in so doing he instantiates himself as the conduit for God's will. As such, he sets himself up as the supreme authority in this world. Thus he takes the seat of worldly ruler and so maintains his own as the highest caste.

[2] The rivalry between the knightly noble and the priest can be seen in §7 and also is hinted at in the preceding mentions of the knights and priests in GMI.

5 Belief

Before fully resolving the contradiction it may have been noted by some close readers of GMIII that Nietzsche does make reference to a sense of 'conviction' (§11) that the priest has in AI: 'This ideal constitutes not only the conviction of the ascetic priest, but also his will, his power, his interest'. But this need not lead one to assert that the priest must therefore believe in the other-worldly aspect of AI proper. Rather, when one considers §20 it is seen that on the contrary, the priest's 'conviction' relates to his application of the ascetic 'balm' in 'good conscience'. That is, in §20 Nietzsche gives the priest a slight reprieve from the ongoing attack on his role in the genealogy of morals (though in fact it is little reprieve at all, for he tells us that the priest's actions have severely retarded the flourishing of mankind). He suggests that the priest honestly believed that furnishing the slave with AI proper could effectively solve the problem of *ressentiment*. §11 also has this conviction stand as only a small part of what the ascetic ideal means to the priest. The much more emphatic following statement has it that the priest's '*right* to exist stands and falls with this ideal'. This relates to how the priest must administer AI to the slave as a means to gain control as was (iii) in our introduction.

6 Conclusions

Now it is possible to make explicit that which should already have become apparent - that which led to the contradiction was an initial misinterpretation of (i) and (ii).

Though it is true that the priest is (i) a creature of *ressentiment*, it does not follow from this that he is slave-like. By grace of his nobility (and unlike the slave) he is able to overcome death alone and so is not in need of AI as a meaning for and a possibility of escape from suffering. But (ii) still holds; he certainly does still require AI in a sense, so in what sense is it required? It is in the sense that AI is the only means through which the priest is able to maintain his status as the highest caste in society (in his role as mediator between God and man). Moreover after having sold the slave AI proper it is only AI proper which lends a reason for his existence.

In response to the question, the priest then, has recourse to both kinds of AI; both AI proper (which as noted, is his own creation) and ascetic procedures (his 'good'). However, the key element to remember is that he never need believe in AI proper for he needn't have it to affirm his life. Rather, he needs

AI proper only so that he can sell it to the slave, and in so doing retain his position in the highest caste and become victor in his war against the knightly nobles. For they must eventually succumb to the whispering of the slaves, and either face suicidal nihilism, or accept the priest's ascetic ideal.

There is a tension that some may feel in the interpretation given here and that is why the noble must eventually face suicidal nihilism and succumb to the priest. I believe that this tension is readily resolvable by considering the noble's intellectual level; however that is a task for another paper.

References

[1] Nietzsche, F. (2008) *On the Genealogy of Morals*. Oxford: Oxford University Press.

[2] Owen, D. (2007) *Nietzsche's Genealogy of Morality*. Montreal: McGill-Queens University Press.

[3] Ridley, A. (1998) *Nietzsche's Conscience: Six Character Studies from the Genealogy*. New York: Cornell University Press.

The Knowledge Argument, Phenomenal Concepts and Social Externalism

Bastian Christofer Stern
University of Cambridge (Trinity College)

Introduction

Jackson [11] asks us to imagine a neuroscientist, Mary, with knowledge of every physical truth. Mary has spent all her life in a black-and-white room so that all the colours she has experienced are black, white and various shades of grey. Crucially, it seems that despite knowledge of all physical information, she gains new factual knowledge upon leaving the room. For example, when she sees a ripe red tomato for the first time, she discovers what it is like to see red. From this Jackson concludes that there must be information over and above physical information, and hence physicalism, understood as the view that all information is physical information, is false. This is Jackson's original formulation of the knowledge argument (henceforth KA).[1]

In response to KA, proponents of what Stoljar [15] has labelled the 'phenomenal concepts strategy' (PCS) aim to explain Mary's lack of knowledge of the phenomenal truths she discovers upon leaving the room by appealing to special features of our phenomenal concepts.

I here understand phenomenal concepts as concepts of phenomenal properties, or alternatively, as the constituents of phenomenal beliefs, i.e. beliefs that attribute phenomenal properties. Phenomenal properties characterise what things are like for a subject. An example of a phenomenal property frequently used for illustration below is phenomenal redness, the property that characterises what it is like to see red.

Derek Ball [3] and Michael Tye [17] (B&T) argue that PCS as a response to KA is undermined because social externalism applies to phenomenal concepts.

Proponents of PCS, as Ball and Tye understand it and as I will understand it below, explain Mary's failure to know the phenomenal truths she later

[1] I believe that KA is better developed in terms of deducibility and necessitation (cf. Chalmers [6]). However, as everything I say below can be applied *mutatis mutandis* to such a formulation, I focus on Jackson's original argument.

discovers with her lack of the relevant phenomenal concepts.[2] Without these concepts, she cannot even entertain the relevant contents, and Mary's lack of phenomenal knowledge whilst inside the room does not cry out for a further explanation that appeals to the existence of information over and above physical information.

Mary's lack of the relevant phenomenal concepts before leaving the room is in turn typically explained by proponents of PCS with appeal to their distinctively strong possession conditions. This is usually taken to be an experiential requirement: To possess these phenomenal concepts, one must have had the experiences to which they apply. (I call concepts that have this requirement 'strong phenomenal concepts'.) Moreover, the proponent of PCS will argue, it is perfectly coherent to claim that these strong phenomenal concepts pick out physical properties, and accordingly, the appeal to strong phenomenal concepts allows him to explain Mary's epistemic situation without having to make serious concessions to the anti-physicalist.

B&T now argue that because social externalism applies to phenomenal concepts, Mary can already, while still in her room, possess the concepts with which she thinks about phenomenal properties after leaving the room. This means that there are no strong phenomenal concepts, which undermines the explanation PCS provides for Mary's progress, and thereby also PCS as a response to KA. Ball further argues that the KA itself relies on Mary acquiring strong phenomenal concepts ([3] pp. 941-3), and is undermined by his arguments against their existence. Here I simply grant Ball this latter point. If, as I go on to argue, B&T's arguments do not force us to deny the existence of strong phenomenal concepts, the central thesis I intend to argue for, that the applicability of social externalism to phenomenal concepts leaves the dialectic over KA unchanged, is secured either way.

After sketching B&T's arguments in Part II, I outline two promising responses to B&T's arguments on behalf of the proponents of KA and PCS in Part III, which rest on a distinction between the belief-contents we would ordinarily ascribe to a subject, and the two-dimensional contents of his beliefs. The upshot will be that B&T's arguments do not affect the dialectic over KA.

Finally, in part IV, I discuss Alter's [2] related line of response to B&T. I

[2] For proponents see e.g. Harman [9], Tye [16], Papineau [13], Hellie [10], Kirk [12]. Note that *pace* B&T, there are versions of PCS for which it is arguably inessential that Mary gains new concepts (cf. Alter [1]).

show that it is inelegant, and that without the 2-D-framework, which renders it superfluous, it lacks the expressive resources to do the work it is meant to do.

1 Ball and Tye's Arguments

Social externalists hold that which concepts one possesses depends constitutively on one's social and linguistic environment. To see how this connects with B&T's argument, first consider social externalism as applied to the familiar example of the concept *arthritis*. To use Burge's [4] famous thought-experiment, imagine a patient who says to his doctor "I have arthritis in my thigh". This patient, despite his radically mistaken conception of arthritis (arthritis only occurs only in the joints), nevertheless seems to possess *arthritis*. This is indicated by patient's acceptance of the doctor's correction that he was in error when he believed that he had arthritis in his thigh, and by the fact that patient and doctor can share e.g. the thought that the patient has arthritis in his ankles.

The thought-experiment suggests that if a subject uses 'arthritis' deferentially, i.e. if he intends to use the expression for the same phenomenon to which the experts in his community apply it, he can possess *arthritis* despite having an incomplete and distorted understanding of what arthritis is, simply in virtue of being part of a linguistic community which has a public expression for arthritis.

B&T argue that analogous considerations apply to concepts like that of phenomenal redness, because despite having an impoverished grasp thereof, someone in a community with the relevant public language expression can think thoughts like (1) What it is like to experience red is not a number (Stoljar [15]), or (2) What it is like to experience red is more similar to what it is like to experience orange than what it is like to experience green (Tye [17] p. 66).

Crucially for B&T, it seems perfectly plausible that Mary could interact with normal perceivers linguistically whilst still in her room, and thereby come to believe (1) and (2). As in the arthritis-case, the explanation is straightforward: She intends to use the term that expresses the relevant phenomenal concept, 'what it is like to experience red', with semantic deference to her interlocutors who have had rich phenomenal experiences. Accordingly, Mary does not have to experience phenomenal redness herself to think thoughts involving the concept of phenomenal redness. Because this reasoning can be extended to any other phenomenal concept, there are no strong phenomenal concepts.

These considerations rely on the expressibility of phenomenal concepts in natural language. One may object that ‘what it is like to see red’ cannot be the correct expression of the relevant phenomenal concept, because the former is complex whilst the latter is simple. However, nothing stops a community from introducing a simple expression ‘R’ for phenomenal redness (Tye [17] p. 69). By being part of this linguistic community, Mary could acquire the simple concept that ‘R’ expresses, and think thoughts involving that concept.

The believer in strong phenomenal concepts could now respond that contrary to first appearances, Mary’s phenomenal beliefs are constituted by different concepts before and after leaving the room. However, B&T argue, this response has various consequences that are hard to accept, as the following considerations are meant to demonstrate:

(a) *Intrapersonal (intertemporal) agreement*
Mary can share thoughts and agree with her former self after leaving the room. For instance, she might have thought that phenomenal redness is not a number, and this thought may be confirmed after leaving the room. This confirmation would be impossible if there were two phenomenal concepts in play here, and two distinct thoughts being thought before and after leaving the room (Ball [3] p. 952; Tye [17] p. 67).

(b) *Interpersonal agreement*
Mary could share the thought and agree with someone outside the room who has had rich phenomenal experiences e.g. that phenomenal redness is not a number. However, this would be inexplicable if they did not possess the same concepts (Ball [3] pp. 952-2; Tye [17] p. 68).

(c) *Too much new knowledge*
Thirdly, Ball ([3] p. 953) objects that despite already having well-developed views about colour experience, if Mary gains a new phenomenal-redness-concept upon leaving the room, she will implausibly gain a large number of new pieces of knowledge corresponding e.g. to (1) and (2), now constituted by the newly acquired strong phenomenal concept.

It is unclear what to make of this last objection. If Ball’s point is that it is implausible that despite already having such beliefs, Mary acquires them again, his opponent, who claims that Mary acquires *new* beliefs involving new concepts, can happily grant this. Accordingly, I only respond to (a) and (b).

2 Addressing the Arguments

I think the best way to resist B&T's arguments is to hold that the relation between the concepts we ascribe in ordinary belief-ascriptions and those that actually constitute people's beliefs is more complex than B&T assume.

The way I develop this idea here is, on the one hand, to concede in the face of B&T's arguments that there may be no difference in the belief-contents we would ordinarily ascribe to Mary before and after leaving the room, but to argue on the other hand that her belief-tokens nevertheless differ in their associated two-dimensional contents. Accordingly, different concepts with different two-dimensional profiles will constitute these beliefs, and the proponents of KA and PCS can hold that concepts with some specific two-dimensional profiles are strong phenomenal concepts: For their possession, Mary must have had the relevant experiences.

I first briefly explain how Chalmers' 2-D-framework can be applied to beliefs and concepts (cf. Chalmers [5]). For a given belief-token B, we can identify intensions, i.e. functions from possible worlds to the truth-value of B at that world. The two-dimensionalist associates beliefs with two kinds of intensions. The primary intension of B is a function that takes worlds considered as actual as its argument, and the truth-value of B at that world as its value. As a heuristic, the value of a primary intension of B given a particular argument, a centred world W centred on the believing subject, is the answer to the question: "If W actually obtains, what is the truth-value of B?" The secondary intension of B is a function that takes worlds considered as counterfactual as its argument, and the truth-value of B at that world as its value. Here, the value of the secondary intension of B given a particular argument, world W*, is the answer to the question: "Had W* obtained, what would have been the truth-value of B?" The primary and secondary intensions for a given concept-token C are defined analogously, with the difference that they map worlds considered as actual and counterfactual respectively to the extension of the concept at that world.

Ordinary belief-ascriptions are sensitive to both the primary and the secondary intension of the subject's belief token (cf. Chalmers [7]). Consider the case in which Peter ascribes Lois Lane a belief by saying "Lois believes that S", e.g. "Lois believes that Clark Kent is muscular." For this ascription to be true, Lois's belief must share the secondary extension of S as uttered by Peter; Lois's belief and S must be true in the same worlds considered as counterfactual.

However, despite ‘Clark Kent is muscular’ and ‘Superman is muscular’ sharing the same secondary intensions, it may be false for Peter, on this occasion or in general, to ascribe Lois the belief that Superman is muscular. Accordingly, a further constraint on belief-ascription is that the subject’s belief has a primary intension appropriate for the ascribed content. For a primary intension to be appropriate for the attribution of *Clark Kent is muscular*, for instance, Lois’s belief has to refer to Clark Kent via a concept with a primary intension that is associated with his role as Clark Kent, e.g. one that picks out whoever is the shy investigative journalist working at the Daily Planet. If it referred to Clark Kent via a concept with a primary intension associated with Superman, e.g. one that picks out whoever is the famous superhero from Krypton, the belief attribution would be false. Crucially for my purposes, however, there will be a whole class of primary intensions that qualify as ‘Clark Kent’-appropriate. Accordingly, ordinarily attributed belief contents are sensitive to primary intensions, but not as fine-grained.

We can distinguish different phenomenal-redness-concepts which come apart on the 2-D-framework, but which ordinary belief-ascription conflates.[3] First consider the concept expressed by the public language term ‘phenomenal redness’. The extension of ‘phenomenally red’ is determined via a relation to things that are red in the external, non-phenomenal sense, as the person who acquires ‘phenomenally red’ learns to apply this expression to the experiences that are typically brought about by objects that are red in the external sense.

At least two relational concepts can be expressed by ‘phenomenally red’, corresponding to taking the relevant relation to be the relation of typically being caused by external red things in the speaker himself, or of typically being caused within normal members of the speaker’s linguistic community. I will call these $phenred_{ind}$ and $phenred_{com}$ respectively.

We also have to make room for a further, what we might call pure concept of phenomenal redness, which I call $phenred_{pure}$. This becomes clear when we consider that when Mary leaves the room, she will not only learn that experiencing red has such-and-such a quality, but also that the quality typically caused by red things, both in herself and in other members of her community, is that very same quality. If we take ‘such-and-such a quality’ to express $phenred_{pure}$, these new pieces of knowledge correspond (roughly) to $phenred_{ind} = phenred_{pure}$, and $phenred_{com} = phenred_{pure}$. As these are clearly cognitively

[3] The first part of this taxonomy follows Chalmers [6].

significant pieces of knowledge, and as cognitive significance of identities involving particular concepts is a plausible test for their distinctness, *phenred*$_{pure}$ is a distinct concept from the other two.

We can give a rough explanation of the differences between these concepts in terms of the 2-D-framework. All three concepts mentioned above rigidly refer to a phenomenal property; they have necessary secondary intensions. When they feature in the beliefs of a non-deviant individual in a non-deviant community, they all (rigidly) refer to phenomenal redness.

However, the primary intensions of these concepts differ, which is why the identities linking them are cognitively significant and a posteriori. The primary intension of *phenred*$_{com}$ picks out the quality typically caused by red objects in members of the community to which the subject at the centre of the world belongs, that of *phenred*$_{ind}$ the quality caused by red objects in the individual at the centre. The primary intension of a pure phenomenal concept, like *phenred*$_{pure}$, picks out the same quality in all worlds.

Now crucially, I claim that all T&B have shown is that we have to make room for a further concept of phenomenal redness, namely a deferential concept which Mary can possess merely in virtue of being part of a linguistic community. I refer to it as '*phenred*$_{def}$', and it can be roughly glossed as *whatever experts in my community refer to when they use 'phenomenal redness'*.

As before, *phenred*$_{def}$ is distinct from *phenred*$_{pure}$ because *the phenomenal quality experts refer to when they use 'phenomenal redness' is such and such a quality*, which corresponds (roughly) to the identity *phenred*$_{def}$ = *phenred*$_{pure}$, is another cognitively significant piece of knowledge Mary gains upon leaving the room. In terms of the 2-D-framework, the primary intension of *phenred*$_{def}$ will be that of *phenred*$_{ind}$ or *phenred*$_{com}$ in worlds in which 'phenomenally red' is used as in ours, but differ completely in social twin-earths in which this expression is used differently.

How does Mary's transition from a piece of knowledge involving *phenred*$_{def}$, for example *phenred*$_{def}$ is instantiated, to the corresponding piece of knowledge involving *phenred*$_{pure}$, in this example *phenred*$_{pure}$ is instantiated, constitute epistemic progress? Chalmers' metaphor of pieces of knowledge as dividing the epistemic space, the space of all epistemically possible scenarios, provides a helpful way to visualise this (cf. Chalmers [8]). Each substantial piece of knowledge partitions the epistemic space by excluding some scenarios as epistemically impossible for the knowing subject, whilst leaving others epistemically possible. It is then natural to say that gaining a piece of knowledge constitutes epistemic progress if more scenarios are excluded as epistemically impossible for the subject than were excluded given his total prior knowledge.

Gaining knowledge of *phenred$_{pure}$ is instantiated* when one previously only knew *phenred$_{def}$ is instantiated* narrows down the scenarios left open. Whilst knowledge of *phenred$_{def}$ is instantiated* leaves open the possibility that phenomenal redness is not instantiated (e.g. because speakers in the subject's community use 'phenomenal redness' to refer to phenomenal greenness), the thought *phenred$_{pure}$ is instantiated* effectively zooms in on only those worlds in which phenomenal redness is instantiated. Accordingly, claiming that Mary acquires the belief *phenred$_{pure}$ is instantiated* upon leaving the room explains her epistemic progress.

This latter consideration may help to support an even better response to B&T's arguments (a) and (b). This consists in saying that whilst Mary gains an *additional* phenomenal concept, she does so without her deferential phenomenal concept and the beliefs that involve it becoming replaced. They continue to co-exist with the beliefs she gains, and do co-exist with these beliefs in subjects outside the room. Accordingly, intra- and interpersonal agreement are even more elegantly explained: The deferential phenomenal redness concept pre-release-Mary shares with others and her future self makes them possible.

B&T (Ball [3] p. 953; Tye [17] p. 67) oppose this move because they regard the claim that we possess two phenomenal redness concepts as introspectively implausible. I concede that perhaps we cannot directly attend to two beliefs involving different phenomenal-redness-concepts. However, good reasons to nevertheless posit their existence stem from thinking about how Mary's beliefs before and after leaving the room, as well as our beliefs, partition the epistemic space into scenarios that are doxastically impossible and those left open. It is very plausible that during a transition like Mary's, no doxastic possibilities would newly open up. Neither does it seem that there are doxastic possibilities open for us that are not open for Mary in the room. As knowledge involving *phenred$_{pure}$* does not close all gaps in the epistemic space which the same knowledge involving *phenred$_{def}$* closes, the best explanation is that Mary does not experience a transition of beliefs, but merely the addition; her belief *phenred$_{def}$ is instantiated* remains when she leaves the room, and can be shared with her future self and those outside the room.

Accordingly, both the proponents of KA and PCS can remain unimpressed by B&T's arguments. They can explain Mary's progress by appealing to a new strong phenomenal concept Mary acquires upon leaving the room, $\textit{phenred}_{\textit{pure}}$. Although this is in some tension with the data B&T adduce (it entails that Mary does not, at the level of two-dimensional content, share thoughts with her future self and those outside the room), they can accommodate this data at the level of ordinary ascribed content. Alternatively, they can plausibly hold that post-release-Mary and those outside the room share with pre-release-Mary beliefs involving $\textit{phenred}_{\textit{def}}$, so that the data is accommodated on both levels. The epistemic advantage of those outside the room consists in also knowing contents involving $\textit{phenred}_{\textit{pure}}$. Accordingly, B&T's arguments leave the dialectic over KA unchanged.

3 Knowledge$_P$ and Knowledge$_M$

Alter [2] suggests a related response to B&T's arguments. He grants B&T that phenomenal concepts do not have strong *possession*-conditions, but holds instead that they have strong *mastery*-conditions, where mastery is understood as non-deferential, or full possession. Proponents of this concept-mastery response can then grant B&T that Mary already has possession of the relevant phenomenal concepts in her room, and phenomenal knowledge under these concepts (which Alter calls "knowledge$_\mathrm{p}$"), but that her epistemic progress is explained by the knowledge under concepts which she masters (knowledge$_\mathrm{m}$), which she can only acquire after leaving the room, as mastery of concepts like that of phenomenal redness requires the relevant experiences.[4]

Whilst my strategy in part III reinstates an experiential condition in the face of B&T's arguments by identifying a level of content at which ordinary belief-ascription partly conflates conceptual differences, Alter can reinstate such a condition by identifying a difference in the relation Mary in which stands to the contents and concepts that constitute her phenomenal beliefs before and after leaving the room, to which ordinary belief- (and knowledge-) ascription is also insensitive.

Like my response, the mastery response suggests that the dialectic over KA remains unchanged. The proponent of KA can claim that Mary's acquisition of new knowledge$_\mathrm{m}$ despite knowledge$_\mathrm{m}$ of the physical truths is in tension with physicalism; the proponent of PCS can explain Mary's failure to ac-

[4] See also Rabin [14].

quire knowledge$_{m}$ while still in the room by pointing to the strong mastery conditions of phenomenal concepts.

However, the response proposed in part 2 is preferable to Alter's mastery response for several reasons. Firstly, by employing the 2-D-framework, we can avoid a worry that Tye ([17] p. 129) expresses about the mastery response, namely that if all that changes when Mary leaves the room is the way she grasps propositions she already knew pre-release, it is difficult to see how she makes the discovery she seems to make.

In contrast, in addition to explaining Mary's epistemic progress in terms of the closing down of epistemic possibilities as outlined above, the two-dimensionalist can take the propositional objects of Mary's knowledge to be two-dimensional entities that involve both primary and secondary intensions (cf. Chalmers [7]). The propositional objects then do change when Mary leaves a room, and her discovery is naturally explained.

In addition, the mastery/possession-distinction is insufficiently fine-grained, so that its proponent may any case be forced to use the 2-D-framework. For instance, consider the difficulties Alter experiences in responding to the possibility that Mary introduces a phenomenal concept by stipulation, e.g. by stipulating 'phenred$_{\text{stip}}$' to refer to whatever it is like for her to see red. It seems that possessing the concept 'phenred$_{\text{stip}}$' expresses, *phenred$_{stip}$*, is one way mastering the concept of phenomenal redness.

The only response open to Alter, apart from using the 2-D-framework, is to declare by fiat that *phenred$_{stip}$* is not a phenomenal concept in the sense relevant to KA ([2] p. 8). Were he to instead help himself to the 2-D-framework, he could point to the relevant difference between *phenred$_{stip}$* and the concept central to KA, *phenred$_{pure}$*, namely the difference in their primary intensions. However, if we uses the expressively richer 2-D-framework, the elegant responses to B&T sketched in part III come for free; the there is no work left to do for the concept/mastery distinction.

4 Conclusion

I conclude that proponents of KA and PCS can remain unimpressed by B&T's argument that because social externalism extends to phenomenal concepts, Mary can possess all beliefs about phenomenal redness that she has after leaving the room whilst still in her room, and that consequently there are no strong phenomenal concepts.

By making use of the 2-D-framework, they can firstly argue that whilst Mary's beliefs may have a kind of content (ordinary ascribed content) that is invariant before and after leaving the room (as B&T's supporting arguments suggest), Mary does acquire beliefs with a new two-dimensional content. They can claim these to be constituted by strong phenomenal concepts. Defensibly, they can do even better and acknowledge that Mary's deferential beliefs about phenomenal redness stay constant, but that she simply acquires additional beliefs about phenomenal redness under a strong phenomenal concept.

Alter's related mastery-response to B&T is less elegant, and without the 2-D-framework, which brings the elegant responses sketched in part 2 with it for free, it lacks expressive power required in the context of the KA. If the 2-D-framework is adopted, there is no work left to do for the mastery-response.[5]

References

[1] Alter, T. (2011) 'Tye's New Take on the Puzzles of Consciousness', in *Analysis*, 71(4): 765-775.

[2] Alter, T. (unpublished) 'Social Externalism and the Knowledge Argument'.

[3] Ball, D. (2009) 'There are No Phenomenal Concepts', in *Mind*, 118: 935-62.

[4] Burge, T. (1979) 'Individualism and the Mental', in *Midwest Studies in Philosophy*, 4: 73-121.

[5] Chalmers, D. J. (2002) 'The Components of Content', in D. J. Chalmers (ed.) *Philosophy of Mind: Classical and Contemporary Readings*. Oxford: Oxford University Press.

[6] Chalmers, D. J. (2004) 'Phenomenal Concepts and the Knowledge Argument', in P. Ludlow, Y. Nagasawa & D. Stoljar (eds.) *There's Something About Mary: Essays on Phenomenal Consciousness and Frank Jackson's Knowledge Argument*. Cambridge, MA: MIT Press.

[7] Chalmers, D. J. (2011) 'Propositions and Attitude Ascriptions: A Fregean Account', in *Noûs*, 45(4): 595-639.

[5] Thanks to Tim Crane for helpful comments.

[8] Chalmers, D. J. (forthcoming) 'The Nature of Epistemic Space', in A. Egan and B. Weatherson (eds.) *Epistemic Modality*. Oxford: Oxford University Press.

[9] Harman, G. (1990) 'The Intrinsic Quality of Experience', in *Philosophical Perspectives*, 4: 31-52.

[10] Hellie, B. (2004) 'Inexpressible Truths and the Allure of the Knowledge Argument', in P. Ludlow, Y. Nagasawa & D. Stoljar (eds.) *There's Something About Mary: Essays on Phenomenal Consciousness and Frank Jackson's Knowledge Argument*. Cambridge, MA: MIT Press.

[11] Jackson, F. (1982) 'Epiphenomenal Qualia', in *Philosophical Quarterly*, 32: 127-36.

[12] Kirk, R. (2005) *Zombies and Consciousness*. New York: Oxford University Press.

[13] Papineau, D. (2002) *Thinking About Consciousness*. Oxford: Oxford University Press.

[14] Rabin, G. (2011) 'Conceptual Mastery and the Knowledge Argument', in *Philosophical Studies*, 154(1): 125-147.

[15] Stoljar, D. (2005) 'Physicalism and Phenomenal Concepts', in *Mind and Language*, 20(2): 296-302.

[16] Tye, M. (2000) *Consciousness, Color, and Content*. Cambridge, MA: MIT Press.

[17] Tye, M. (2009) *Consciousness Revisited: Materialism Without Phenomenal Concepts*. Cambridge, MA: MIT Press.

Professional Wrestling and Reality Modification: The Impossibility of Self-Deception

Samuel Thomas
University of Leeds

Introduction

The notion of self-deception is intriguingly paradoxical and notoriously difficult to define. Nevertheless, when one mentions the concept people generally know what it means and can describe instances of it.

In this essay, I will draw attention to the paradox of self-deception and will go on to argue that truly deceiving one's self is categorically impossible. I will subsequently attempt to offer an alternative explanation for the phenomenon, which will revolve around augmenting the traditional interpersonal model by placing greater emphasis on the motivations of the deceiver. I will go on to argue that so-called self-deception really amounts to nothing more than intentional and conscious reality modification, in which no real deception takes place and that the motivation behind self-deception is related to the general happiness of the deceiver. I will then offer a conciliatory defence of this potentially controversial theory by way of familiar and less controversial examples.

1 The Paradox

In its traditional conception, self-deception is paradoxical and modelled on interpersonal deception i.e. deception by another. Roughly speaking, self-deception is just that: it is the act of deceiving oneself into thinking that a truth value of a given proposition is the opposite of what it in fact is. How can I believe that I have done enough studying while simultaneously not believing that I have done enough? I cannot. In other words, it seems that the self-deceiver would have to hold contradictory beliefs, which seems implausible. Therefore self-deception is impossible. Indeed, I do not wish to dispute this claim. Self-deception *simpliciter*, in its traditional and strictest logical form, is certainly impossible. Positing that an agent can truly deceive oneself leads to a paradox in which the self-deceiver ends up consciously believing both p and not-p.[1] This paradox alone is sufficient to cast doubt upon just what it is we think is going on in cases of self-deception. Working on the assumption that we do not wish to give up on there being such a thing in the first place, what remains is to account for the phenomena in a different way. My intention is to salvage and adopt the interpersonal model while maintaining that cases of self-deception involve no actual deception. I will attempt to do this by appealing to the intentions of the deceiver in such cases and placing a greater importance on the motives underlying them.

In the interpersonal case, the deceiver seeks to alter the set of beliefs of their victim (in spite of the actuality) in order to instigate what they judge to be a more preferable state of affairs that would otherwise fail to transpire. Examples of these states of affairs include things as elaborate and serious as material gain or indeed things as relatively small and simple as the mild amusement of the deceiver. Indeed people deceive each other for various reasons to varying degrees. Fundamentally though, people deceive others because of some perceived gain they believe would result from them doing so.

Analogously, in the self-deception case, I would argue that the deceiver also seeks to instigate a more preferable state of affairs. However, given they already know the truth, and are unable to contradict their beliefs to render the preferable actual, all that remains is to alter their behaviour and thereby behave as if they believed the more preferable state of affairs was real. In other words, just as the interpersonal deceiver modifies their victims set of beliefs because they judge that it is in their interest to do so, so the self-deceiver alters their own behaviour to match the false belief because they think it is their interest to do so. In either case though, it is not the belief itself that is important; it is the behaviour that holding that belief would elicit which is the driving force behind the act of deception. Put simply, one could imagine the interpersonal deceiver deliberating: "I believe that p, but I believe it is in my interest that you believe and thus behave as if not-p", where they are motivated to alter your beliefs in order to alter your behaviour. Likewise, the self-deceiver might say: "I believe that p, but I believe it is in my interest that I behave as if I believe not-p" where they are motivated to ignore the actual in favour of the preferable. For example, the doting husband may 'deceive' himself into thinking that his wife is being faithful despite overwhelming empirical evidence to the contrary, for the simple fact that the not-p reality in which his wife is faithful is more palatable than the lamentable though

[1] Deweese-Boyd [2] §1.

actual p reality in which she is not.

I will admit that on the face of it, the idea of making the conscious, rational choice to modify ones reality in order to make life more palatable sounds peculiar to say the least, but it is not quite as alien as one might think. To try and clarify this idea, I shall take this opportunity to invoke a particular aspect of a subject that I have until now been unable to incorporate into any philosophical discourse throughout my time as an undergraduate. That subject is professional wrestling and the particularly relevant aspect is the 'suspension of disbelief'. I would say that wrestling is a guilty pleasure of mine, but I would be lying.

Now, I know professional wrestling is not real, indeed everyone but the youngest of fans knows this. I know that the elaborate theatrics, production values and the dichotomy of Babyfaces (goodies) verses Heels (baddies) is totally staged and make believe. I know that when the wrestler throws a punch, nine times out of ten his fist barely brushes the skin of his opponent and I know that most of the time, rivalries cease to obtain when the stage lights fade to black. But my knowledge of these things does not stop me from identifying with the wrestlers, from cheering for the good guys and booing the bad guys, or from marvelling at the acrobatics and athleticism. It does not detract from my overall enjoyment of professional wrestling as a spectacle and as an experience, for the very simple reason that for the duration of the show; I choose to behave as if I believe that the wrestling world is as real and as genuine as any aspect of the real world I inhabit on a day-to-day basis. In this sense, I *suspend my disbelief* in exchange for a more appealing and enjoyable experience overall.

The same basic idea can be applied to literature, films, soap operas and indeed any medium which appeals to a fictional world. For example, when a film moves us to tears, it is the suspension of our disbelief that renders it possible. The characters do not exist, their respective plights are mere plot points in a screenplay but as soon as the opening scene unfolds, we willingly immerse ourselves in their world. That is to say, we modify our own reality in order to identify and interact with the characters and their situations and experiences, however emotional, fantastical or amusing they may be. Indeed, such is the beauty of fiction. These unashamedly self-indulgent examples nevertheless illustrate a philosophical point with regard to our current concern namely, that we regularly adopt alternative versions of reality and indeed that it can have some positive uses. So in the case of the doting husband, he essentially chooses to suspend his disbelief regarding his wife's infidelities in favour of a more palatable alternative: that she has been faithful and that everything is all right. In other words, instances of self-deception are not so much cases of convincing oneself that a particular truth value or set thereof is the opposite of what it in fact is. Rather, it is the act of choosing to disregard that truth value, or in other words to modify one's reality as to that particular truth value by weighing up the implications that adopting it would have on one's immediate psychological well-being or happiness.

2 Some Objections

There are two kinds of potential objections to this account of self-deception that I wish to address. The first is the foreseeable hostility towards 'reality modification' and the second is my presumption of conscious intention.

For clarity; I have argued that cases of so-called self-deception can be more accurately described as being instances whereby an agent:

(1) Believes that p.
(2) Judges that not-p is a more desirable state of affairs than p.
(3) And therefore behaves as if not-p.

The first and most specific objection is that suspending ones disbelief to better enjoy a work of fiction is incomparable to modifying ones reality and living a lie in real life. Admittedly there are some *prima facie* differences. The implications of living a lie are far greater given that you are interacting with a real and tangible world. Moreover, the timescales of the respective suspensions of disbelief are different insofar as a film (for example) has a predetermined length but self-deception could potentially last a lifetime. I would argue that objections of this kind stem from a misguided though perfectly understandable apprehensiveness concerning the wider implications that adopting such a theory could potentially entail. To me it seems these differences are cosmetic and the distinctions arbitrary.

More fundamentally however, one might object to self-deception being a conscious and intentional act in the first place.[2] In terms of self-deception being a conscious act, we need only posit a level of self-awareness on par with something more familiar and mundane. For example, when I walk, I do so consciously, but I need not focus on each successive step in order to continue

[2] Bermúdez [1].

walking. Similarly with fiction, the initial conscious act of acceptance is sufficient; the rest follows as you are absorbed into it. Thoughts of "I am immersed in a fiction" scarcely enter the mind, thus accounting for the apparent lack of immediate awareness of self-deception.

With regard to intention, I would maintain that deception of the self entails some sort of intention for otherwise you are presumably being deceived by your sub-conscious or some other agency-depriving entity. At any rate self-deception without conscious choice amounts to self-deception without the 'self' which is for all intents and purposes the same as interpersonal deception.

Indeed, I think that arguing for non-intentional self-deception opens a proverbial can of worms with regard to culpability and mental illness. For if we have no control over the phenomenon whatsoever, this would suggest that any and all cases of self-deception are the result of some faulty-thinking or mental disorder which is something I think few would be willing to accept. For this reason, I discount cases of psychosis in which the individual's perceived reality is in its entirety, totally inconsistent with the real world due primarily to their agency and faculties of reason not being their own. The point at which ones modified reality becomes totally inconsistent and unworkable given the actual world is the point at which self-deception descends into psychosis. I discount cases where an agent is not in command of all the relevant facts and faculties required to make a reasoned judgement. Indeed, such cases fall under the category of self-deception by error which as most would agree should be cast aside. Similarly, I would also place cases of 'twisted self-deception'[3] as being in this category. For example, if our doting husband was instead extremely jealous and believed his wife was being unfaithful despite overwhelming empirical evidence to the contrary, this is probably due to some cognitive fault beyond his control for presumably (given the choice) he would choose to acknowledge the more preferable and in this case 'real' truth. So in essence, objecting to self-deception being an intentional and conscious act is to object to the interpersonal model in general, which is another debate in itself and one which I cannot discuss in detail here.

[3] Mele [3].

3 Conclusion

Can you know that p but simultaneously know and believe that not-p? Certainly not. However it is perfectly consistent to believe that p but behave as if not-p. It is this consistency which leads me to conclude that cases of self-deception are devoid of deception. Instead, an agent is fully aware of what is going on and makes the conscious choice to adopt a particular and ultimately more palatable reality over a less desirable but 'real' truth. To conclude then I have maintained that self-deception is paradoxical and impossible because one cannot hold contradictory beliefs. I have attempted to offer an account of the phenomena by appealing to reality modification on the part of the agent, in which no deception takes place. I have drawn parallels between self-deception and the suspension of disbelief. When people watch films or soap operas; read fiction or 'deceive' themselves; or indeed watch wrestling... they know none of it is real but they engage with it as if it is, in exchange for a more palatable experience overall. In short, 'self-deception' *per se* is a logically impossible misnomer which really describes a phenomenon akin to interpersonal deception and is similar to the suspension of disbelief (which people engage in on a regular basis anyway). In essence, it is the activity of modifying ones reality for one's own ends.

References

[1] Bermúdez, J. L. (2000) 'Self-deception, Intentions, and Contradictory Beliefs', in *Analysis*, 60: 309-319.

[2] Deweese-Boyd, I. (2010) 'Self-deception', in Zalta, E. (ed.) *The Stanford Encyclopedia of Philosophy* [http://plato.stanford.edu/archives/fall2010/entries/self-deception/].

[3] Mele, A. R. (2010) 'Twisted Self-Deception', in *Philosophical Psychology*, 12: 117-137.

Interviews

Timothy Williamson[*]
Oxford University

What is philosophy?

Philosophy is the sort of inquiry in which one asks questions like "What is philosophy?" That answer may look circular: how can it help someone who doesn't already know what the word "philosophy" means? But it is more informative than it may seem, because the same does not apply to other disciplines. For example, physics is not the sort of inquiry in which one asks questions like "What is physics?" Although that is a question about physics that physicists might occasionally ask themselves in an idle moment, it is not a question in physics. The methods of physics, such as experimentation and calculation, are of little help in answering it. The methods of philosophy are of more help there.

When one tries to define philosophy more rigorously, by stating general, precise, non-circular necessary and sufficient conditions for something to be philosophy, one runs into trouble. Just try it. With a little ingenuity, one can find counterexamples to the proposed definition, either things it fits that are not philosophy or things it does not fit that are philosophy. That is not a peculiarity of philosophy. Other disciplines are hard to define rigorously too. Why is it so hard? Over the course of human history, different traditions of inquiry have developed, different intellectual communities with their own questions and ways of answering them, and increasingly their own institutional identities in university departments, journals and the like. We can call them "disciplines". We learn something about these disciplines from school, the internet, and so on, and acquire some capacity to recognize which of them a particular bit of intellectual activity belongs to. Like other recognitional capacities, this is based on similarities to examples already classified. We are not applying a definition secretly written in our heads. Indeed, there is no deep gulf between philosophy and other disciplines. Philosophical logic shades into mathematics, metaphysics into physics, philosophy of mind into

[*] Professor Williamson delivered the keynote paper at the BUPS 2012 Summer Conference on 2-3 June at the University of Leeds.

psychology, discussion of past philosophical theories into history, and so on. Often, when philosophy shades into another discipline, there are significant areas of overlap.

If you want to help someone who has never heard of philosophy recognize it on sight, you could say that it tends towards the general rather than the particular, the essential rather than the accidental, the abstract rather than the concrete, reflection rather than perception, and prose rather than graphs, but that isn't a serious attempt at a definition.

Another way of approaching the question is to ask what a society without philosophy would be missing. For a start, it isn't clear that there could be a fully human society totally without philosophy. Even very young children ask incipiently philosophical questions. It's been said that all children go through a phase of asking "Why?" about everything, and philosophers are the ones who never grow out of that phase. But philosophy does take much more developed and systematic forms in some societies than in others. We can ask what a society is missing if it has philosophy only in less developed and systematic forms. For example, what difference would it make if all philosophy departments in British universities were abolished? Politicians would still appeal to moral principles to justify their policies, and other politicians would still question the relevance or truth of those principles (at least, while democracy lasts). But the place where the consequences of moral principles are most systematically and carefully drawn out, and compared with the consequences of other moral principles, is moral philosophy. It is also the place where the question where the question "Should we live by moral principles at all, rather than by particular moral judgments or perhaps without morality at all?" is most thoroughly considered. Similarly, all those who used mathematics in a society without philosophy would still implicitly rely on mathematical axioms, typically those of set theory. But the place where the question "What entitles us to rely on the axioms?" is most extensively discussed is the philosophy of mathematics, not mathematics itself. A society can survive without systematic critical analysis of its underlying assumptions. Nevertheless, it exhibits a sort of shallowness that makes it, in that respect at least, not the best sort of society in which humans can live. In the long run, moreover, the limitations on its powers of self-criticism may make it inflexible and even lead to its destruction. Philosophy provides the intellectual framework for an especially searching form of self-criticism.

What do you see as the goal of philosophy?

The goal of philosophy is to increase knowledge of the answers to philosophical questions, questions of the kind I have already illustrated. I admit that the philosophical knowledge we have gained so far is extremely partial, although not negligible. Intellectual despair is a cheap response.

Some people say that the goal of philosophy is understanding rather than knowledge (this is said of many other disciplines too). I don't think they have thought enough about the relation between knowledge and understanding. If you don't know why the sky is blue, you don't understand why it is blue either. Similarly, if you don't know why injustice is bad, you don't understand why it is bad. Conversely, if you know why injustice is bad, you are at least a long way towards understanding why it is bad. Maybe you know that some fact explains why injustice is bad without knowing why the fact explains it: if so, what you lack is still knowledge.

By what method do you think that this goal can be achieved?

Philosophy employs many methods, not just one. They are not fundamentally different from those of other disciplines, although philosophy uses them in distinctive proportions. Philosophers formulate and compare theories about the matters that interest them. The standards by which we compare them are similar to those in other areas of science (not just natural science — mathematics is a science but not a natural science). A good theory is strong, simple, elegant, and consistent with what we already know. Exactly why aesthetic qualities such as simplicity and elegance should be a guide to truth is itself a philosophical problem, still largely unsolved, but we rely on them all over natural science, for example in extrapolating from a finite number of data points to a curve. The extent of what we already know is often highly controversial, but such controversy occurs in other sciences too. Philosophers spend lots of time clarifying their theories, to make it easier to tell what follows from them and what does not.

Philosophers rarely do real-life experiments of their own. Sometimes they devise thought experiments, which are really just ways of testing philosophical theories by what they imply about what would be the case in various hypothetical circumstances. We evaluate those implications using the normal human capacity to assess counterfactual conditionals of the form 'If X were the case, Y would be the case'. Self-described 'experimental philosophers' disdain that method as reliance on unreliable 'intuitions', but their critique is undermined by confusion as to what counts as 'intuition'. They think philosophers should do more real-life experiments to find out what different groups of people really think. Philosophy may have something to learn from such opinion polls, if they are properly conducted — preferably by psychologists, who have far more know-how than philosophers about designing and conducting such experiments. The results may well teach us something about how we (philosophers as well as non-philosophers) are inclined to think about various philosophical topics, such as knowledge and moral responsibility. But their bearing on the main issues will in any case be very indirect. The main philosophical questions concern knowledge and moral responsibility themselves, rather than what we happen to believe about them. 'Experimental philosophy' has very little to say constructively about how we should address major philosophical topics.

Of course, as a test of philosophical theories, consistency with what we already know includes consistency with what we already know by experiment and observation, in particular from natural science. If a philosophical theory is inconsistent with known physical facts, it is false. Equally, a philosophical theory is false if it is inconsistent with known historical facts. By a sensible division of intellectual labour, it is usually best to leave physicists and other natural scientists to ascertain the relevant physical facts, historians to ascertain the relevant historical facts, and so on. Many philosophical theories are false because they are inconsistent with quite everyday knowledge, for instance ones that imply that there are no sticks and stones.

A philosophical training puts one in a better position to apply various formal methods. When we can formalize a philosophical theory in a formal language whose logic is well understood, we are in a much stronger to determine what its logical consequences are. Even if we cannot fully formalize it, we may be able to construct idealized formal models of its application to a specific type of situation, in the manner of the natural and social sciences, and thereby learn much about its workings. Formal model-building strikes me as a somewhat underdeveloped aspect of philosophical methodology. That does not make formal methodology a panacea for philosophy. As in every other area of science, applying formal methods mechanically, without good intellectual judgement and a feel for the subject matter, leads to disaster. Much of philosophy will always be done in natural language.

Why do you do philosophy?

I enjoy it, and fortunately I am paid to do something I enjoy.

What do you do?

I teach and write. Most of the people I teach will go on to become professional philosophers, so they will teach and write too. Perhaps you are asking which parts of philosophy I do. I have mostly worked in philosophical logic, philosophy of language, epistemology, and metaphysics. I have published four books — *Identity and Discrimination* (1990), *Vagueness* (1994), *Knowledge and its Limits* (2000), and *The Philosophy of Philosophy* (2007) — as well as quite a few articles. I am currently completing another book, *Modal Logic as Metaphysics*, to be published next year by Oxford University Press. It asks, for instance, whether you could have been absolutely nothing.

Why is what you do important?

Both accepting and rejecting the presupposition of the question both sound bad, in different ways. The philosopher J.L. Austin used to say 'Importance isn't important'. Anyway, I will say something about the importance of the questions that interest me. We might distinguish their practical importance and their theoretical importance. Like most theorists, I am motivated by the latter, even though some of the questions have practical importance too. For example, if you want to programme a computer to engage properly with everyday speech, you must enable it to deal with the vagueness of most of what we say (as when we answer "A little" to the question "Does it hurt?"), and to do that you must have some sort of theoretical model of vagueness. Most theories of vagueness hold that such a model must depart somehow from the dichotomy of truth and falsity. In my book Vagueness, I argue that we can retain the dichotomy and model vagueness as ignorance of hidden sharp boundaries. Questions about the structural limits of human knowledge, such as I discussed in Knowledge and its Limits, have an obvious interest for reflective participants in any sort of general inquiry. The book I'm currently finishing concerns the correct logic for reasoning about the relations between possibility, necessity, and existence. I don't expect everyone else to want to engage in such inquiries themselves, but I hope that they can at least recognize how fundamental they are.

What advice can you give to students in their first few years of philosophical study?

Enjoy philosophy, and discuss it as much as you can with others who enjoy it too. A subtle distinction can be savoured as much as the best food you ever tasted, and lasts longer. If you like games, philosophical debate can be played as the most mind-bending of games, since the rules are at stake too. If you prefer a more serious approach, remind yourself that it is all ultimately in the service of truth (and that the nature of truth is itself a philosophical question). However clever you are, never think you know it all. The best philosophers are often those who keep learning for longest, long after their first few years of philosophical study. When you think you have unmasked an illusion, that too may be an illusion; you may unmask it tomorrow.

Don't treat anyone (living or dead) as a guru. The more they invite such treatment, the less they deserve it. All philosophers make mistakes — though hardly any of them are idiots. To follow and participate in a philosophical discussion, live or on the page, you should be able to see the issues from each side's point of view. Never rely for your understanding of one side on the other's characterization of it. But it also helps if you care very much which side is right, so that every move feels like a threat or an opportunity. That way you are on the alert.

Boring virtues though they may sound, hard work, stamina, and persistence through adversity are as essential in philosophy as they are in every other human activity. A brilliant flash of insight is worth little if it is not properly followed up. 'The devil is in the detail' in philosophy too. If you keep tugging at what looks like a trivial loose end, you may unravel a whole theory (but you may instead find that it really was a trivial loose end). Don't worry about whether what you are saying is clever or original or deep, worry about whether it is true and relevant to the question. Like happiness, those other qualities are best not pursued directly.

Thank you very much!

Kit Fine
New York University

What is philosophy?

Philosophy has no particular subject-matter and no particular method. It is what is left when other forms of rational inquiry appear to give out.

What do you see as the goal of philosophy?

I would like to think it was truth but may be all we can sensibly aim for is clarity.

By what method do you think that this goal can be achieved?

I believe it is important to trust one's instincts. But instincts are like a wild horse. They can take you far but need a firm hand.

Why do you do philosophy?

Because I enjoy it and am puzzled by its questions.

What do you do?

Metaphysics, philosophy of language, logic.

Why is what you do important?

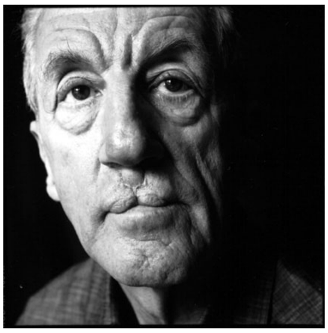
It is important only in so far as philosophy is important and philosophy is important only in so far as its questions and methods of inquiry are important.

What advice can you give to students in their first few years of philosophical study?

If you think the questions of philosophy have easy answers, then you have not understood them.

Thank you very much!

Barry Stroud[*]
University of California, Berkeley



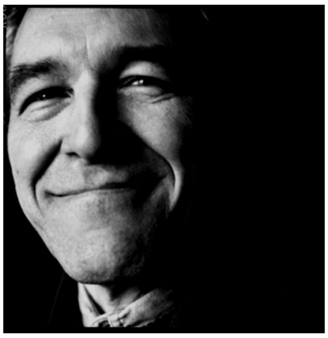
What advice can you give to students in their first few years of philosophical study?

Beginning students should try to read more slowly and more carefully than they are used to reading in other subjects, and they should keep asking themselves what exactly is the question that the text they are reading is trying to answer. The special, distinctive character of typically philosophical questions is what I think it is hard to recognize and appreciate.

Thank you very much!

[*]Professor Stroud was unable to take part in the full interview, but very kindly answered the final question.

Ernest Sosa
Rutgers



What is philosophy?

Philosophy is no one thing; it is the most variegated of disciplines, united only by diverse family resemblances. There is and has been philosophy as way of life; philosophy as therapy, intellectual and practical; and philosophy as a way to approach questions of public policy in various domains to which the philosopher can apply her distinctions and her distinctive understanding. We can contribute if only by placing issues in an illuminating setting or in historical perspective, or by defending unfamiliar options. Philosophers have played that sort of role not only within academic settings but also as public intellectuals, although writing for the general public requires special talents and skills. Philosophers who do this well can make a distinctive contribution. Turning to the traditional core of the discipline, we find the more technical reaches of the various philosophical meta-disciplines, such as philosophy of mathematics, of physics, and many others. And we find the traditional core disciplines of epistemology, ethics, logic, and metaphysics. Thriving international communities of historians also try to illuminate the development of our discipline and the thought of its main contributors through the centuries. Finally, philosophy is variegated not only in respect of topics but also in respect of methods. Formal methods are used in logic, of course, but also in formal epistemology, for example, and in some of the technical reaches of the philosophy of science. Armchair theorizing thrives across the field, and these days there is also experimental philosophy or X-Phi, which brings social-scientific methods to bear. So far the positive contributions made by use of such methods have been modest, both because philosophers have not been well equipped to carry out such research, and because the most resounding contributions have been critical rather than constructive. There is also a gap between (a) the results that one might attain through the use of such social scientific methods, results that would pertain in the first instance to how humans think and speak, and (b) questions about justice itself, or truth, or freedom, or knowledge, etc., the subject matter that has concerned philosophy through the ages.

What do you see as the goal of philosophy?

Obviously there is no single goal. Philosophy has always favoured breadth.

It has even spawned many other disciplines. As my answers have already suggested, philosophy has hosted a great diversity of goals, and still does even now, including reflection on how to live, which encompasses reflection on what gives life meaning. Along with such questions pertaining to wisdom and the good life, at another extreme there is the work in many and varied formal disciplines. And there are the questions I have pointed to concerning knowledge, truth, beauty, justice, happiness, wisdom, etc.—the perennial objects of the philosopher's desire to understand.

By what method do you think that this goal can be achieved?

The methods are many and varied, as befits the diversity of topics and objectives.

Why do you do philosophy?

I do it in pursuit of answers for questions that I find gripping. My own work has been mostly in the core disciplines, mainly in epistemology, though also in metaphysics and some in ethics. I've also been very much interested in several historical figures, though my strongest interest over many years has been in Descartes and his epistemological writings, and in the theoreticians of the flourishing proper to humans, going back to the great insights of Aristotle.

What do you do?

In my work in epistemology, I have developed an account of human knowledge in terms of performance normativity. This is a normativity constitutive of knowledge, which explains its special value. On this account, a belief aimed at truth is a cognitive performance, and is to be assessed in the way of such performances generally. Its value is the value of a performance whose success manifests the performer's competence. That is the nature of our most basic knowledge, and it also explains the special value of such knowledge. Knowledge is always "better" than would be the corresponding merely true belief because performance whose success manifests the performer's competence is always thus "better" than performance whose success does not manifest such competence, either because it does not succeed, or because it succeeds through luck as opposed to competence. I take this to address perennial problems of the nature and value of knowledge that appear already in Plato's Theaetetus and Meno.

Why is what you do important?

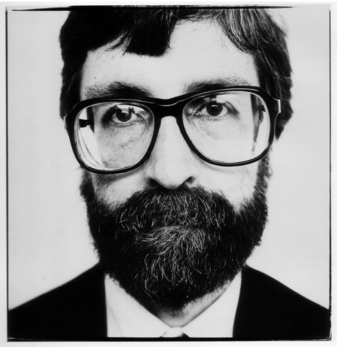
We are rational animals. In the little time we are allotted, once our survival is assured, it is natural for us to wonder. Time is limited and we don't all wonder about the same things, but there are many things worthy of wonder, which repay reflection with lasting and fulfilling insight. They are also worthy of discussion with others capable of appreciating the questions and of enjoying the search for answers. Young people in advanced and emerging societies are allowed a period when they can have the freedom to pursue some such questions, and this is a wonderful opportunity that offers a precious value, quite apart from whatever benefits it may bring in enhanced acuity and intellectual skills.

What advice can you give to students in their first few years of philosophical study?

If you are lucky enough to find some absorbing issues, and to have the time and interest, you should pursue them with zest. Discussion with others similarly minded is a great aid to progress in philosophy, as we see already with the dialogues of Plato, and the thriving of the ancient schools. Of course, there is absolutely no substitute for much reading and reflection on your own, but discussion is also enormously helpful, whether in the classroom or in less formal settings.

Thank you very much!

John McDowell
University of Pittsburg



What is philosophy?

I don't know how to define it except by example. It's the kind of thing that was done by Plato, Aristotle, ..., Aquinas, ..., Descartes, ..., Locke, ..., Kant, ..., Wittgenstein, ... (Different people might cite different paradigms.) My examples belong to a Western tradition, but there's a recognizable sense in which some ancient Indian thinkers, say, were doing the same kind of thing.

What do you see as the goal of philosophy?

To show the fly the way out of the fly-bottle. If you want something grander: to overcome various kinds of obstacles in the way of a healthy understanding of ourselves and our relation to the world.

By what method do you think that this goal can be achieved?

I'm doubtful that the idea of a method is much use in this context. Whatever helps.

Why do you do philosophy?

Because I'm gripped by some of its questions, sometimes to the point of obsession, and I'm lucky enough to be able to earn a living doing it.

What do you do?

Teaching; other kinds of talking, for instance to colleagues; reading; writing.

Why is what you do important?

I think it's obvious that articulating and protecting a healthy understanding of ourselves and our relation to the world is important. (See my answer to question 2.) Similarly with continuing, as self-consciously as possible, a longstanding tradition of intellectual activity. (See my answer to question 1.) It's for that kind of reason that it's important that some people go on doing philosophy. I would be reluctant to claim much importance for my efforts in particular; I try to do my part in what I think of as a communal activity.

What advice can you give to students in their first few years of philosophical study?

Try not to lose sight of the big picture. Otherwise I don't think advice to students of philosophy should be much different from advice to students in any discipline.

Thank you very much!

Issue 5(1)
Summer Conference 2012

Journal of the British Undergraduate Philosophy Society

(Print) ISSN 1748-9393
(Online) ISSN 2051-5359