철학적분석

제 31 호

2014 겨울

한국분석철학회

이 학술지는 한국연구재단의 등재학술지입니다.
이 학술지는 2014년도 (재)플라톤 아카데미의 지원을 받아 출판되었음.
The present academic journal is supported by the Foundation Academia.

철학적분석

제 31 호

차 례

논문

서평

부록

철학적 분석 31 (2014) pp.1-26

Morality Grounds Personal Identity

Bradley Monton

There is a connection between moral facts and personal identity facts: morality grounds personal identity. If, for example, old Sally enters a teletransporter, and new Sally emerges, the fundamental question to ask is: is new Sally morally responsible for actions (and omissions) of old Sally? If the moral facts are such that she is morally responsible, then Sally persisted through the teletransporter event, and if not, Sally ceased to exist.

Keywords personal identity, meta-personal identity, grounding, mereological composition, person, conventionalism

1. Morality and Metaidentity

This is an essay in *meta-personal identity*. Just as, in ethics, we distinguish between applied ethics, normative ethics, and metaethics, so we can make similar distinctions in the field of personal identity:

> *Applied personal identity*: did this particular person who just entered this particular teletransporter just die, or does she continue to exist, over at the exit end of the teletransporter?
>
> *Personal identity*: Is it just psychological continuity and connectedness that matters for persistence of persons, or is it just physical continuity, or is it some combination thereof, or is it persistence of souls, or is it something else?
>
> *Meta-personal identity*: Is there an objective fact of the matter about whether a person does or does not persist, or is it somehow subjective or relative or conventional? If there are objective facts about persistence of persons, what (if anything) *grounds* those facts? If there are no such objective facts, how should we understand everyday utterances that seem to make presuppositions about personal identity (as in: "Fred went home, but he'll be back")?

"Meta-personal identity" doesn't roll off the tongue, alas, so I'll call this "metaidentity" for short. We have to remember, though, that we're not talking about the standard sort of identity (as in: Superman is Clark Kent). Instead, we're talking about *personal identity*, which (by my lights, at least) isn't really about *identity* at all. What personal identity is about is *persistence conditions*: under what conditions does this person survive, and hence persist through time, and under what conditions does this person cease to exist, as a person?

Suppose that Sally undergoes a standard personal identity thought-experiment: brain transplant, teletransportation, memory erasure, ship of Theseus with neurons, what have you. Call “old Sally” the person before she undergoes the experiment, and “new Sally” the person who emerges. (Later we’ll take up cases where more than one person emerges.) Here are two types of questions we can ask:

> (1) Is new Sally the same person as old Sally? In other words, did Sally survive the experiment, or did she cease to exist, as a person?
> (2) Suppose that old Sally did something morally wrong just before the experiment. Is new Sally morally responsible for what old Sally did? (Or, suppose old Sally did something morally right just before the experiment; is new Sally morally commendable?)

Part of the point of this paper is to argue that there is a close connection between these two questions. But which way does the connection go? One could maintain that the personal identity question is the fundamental one, and that the answer to that question will guide us in attributing moral responsibility. This is the dominant view in the literature. Here for example, is David Shoemaker endorsing that view:

> In order to attribute moral responsibility to someone for an act, we must be able to determine that that person is the same person as the person who performed the act. ⋯ What is needed first is a plausible metaphysical account of persons and personal identity to which an ethical theory might then conform and apply. (Shoemaker 1999, 183)

For Shoemaker, the metaphysical account of personal identity is, conceptually, what comes first – the personal identity facts then get inputted into the ethical theory.

On the other hand, one could maintain that the moral responsibility question is the fundamental one, and the answer to that question grounds the facts about personal identity (where the grounding relation is understood in the now-standard way elucidated by for example Gideon Rosen (2010)). This is the view I hold; this is what I'll be arguing for in this paper.

2. Three Assumptions, and an Argument

Let's start with three big assumptions. These are assumptions that I believe, but I won't argue for them here. In this section I'll give my argument for the claim that morality grounds personal identity using these assumptions, and later I'll talk about how my position would change when these assumptions are dropped.

The first assumption is that there are objective moral truths, independent of what we believe about morality. Though it doesn't really matter for the purposes of this paper, the sort of moral realism I endorse is non-natural and non-theistic – the moral truths are not reducible to natural facts and properties, and are not true as a result of the existence (and perhaps choices) of some divine being. Instead, moral truths have the status of other objective truths that are (contentiously) non-natural: modal truths, mathematical truths, and logical truths. This sort of moral realism has been argued for well by for example Michael Huemer (2005) and Erik Wielenberg (2009).

The second and third assumptions naturally go together. The second assumption is that the perdurance theory of persistence is true: objects persist by having temporal parts. A persisting object (in standard spacetime, at least) is a four-dimensional spacetime worm, and this four-dimensional object is composed of instantaneous temporal parts, where each instantaneous temporal part is the object at a particular time (modulo time travel stories). (Some philosophers distinguish between the "stage view" and the "worm view", where the stage view holds that a person is a person-stage, existing at a particular time, while the worm view holds that a person is the whole temporally extended spacetime worm. I don't think there's a metaphysical difference between these two views; sometimes, we refer to person-stages ("Jill is sitting"), while other times, we refer to the whole worm ("Jack had a wonderful life").)

The third assumption is that the doctrine of unrestricted mereological composition is true. For every collection of objects, there is a whole that is composed of those objects as parts. So, for example, there is the 24-hour-long four-dimensional spacetime worm of Obama on July 1 2011, which can be combined with the 24-hour-long four-dimensional spacetime worm of Cheney on January 1 2012, to form a particular object, Obacheny, spread in space and time in a complex, disjoint way.

Given the latter two assumptions, one might think that there is no interesting debate about personal identity to be had. When faced with a personal identity thought experiment, there is an object that exists both before and after the experiment (the mereological sum of old Sally and new Sally, for example), and there is an object that ceases to exist at the time of the experiment (old Sally, for example), and there are an infinite number of other objects (the mereological sum of old Sally, the first two minutes of new Sally, and Obacheny, for example). All these objects exist, according to the doctrine of unrestricted mereological composition, and so it's prima facie unclear how to answer the question of whether Sally persists through the experiment – which of the infinite number of mereological sums of objects are we talking about when we talk about Sally? I maintain that there is a definite answer, and the objective moral facts determine what the definite answer is.

I'll give my whole argument now, and then provide some elaboration. The first step in my argument is to point out that we're talking about *personal* identity, so we're talking about the persistence of *persons*. Suppose that John goes into a permanent vegetative state. John persists as a biological organism, but John (arguably) ceases to exist as a person. Suppose that John later dies and is buried. We make everyday utterances like "John is buried here", and I (perhaps controversially) believe that such utterances are literally true (on a certain interpretation, at least). John does not persist as a person, and does not persist as a biological organism, but does persist as a mereological sum of particles that composed the biological organism pre-death. Similarly, "we are stardust" is also a literally true claim, given a certain interpretation. There are of course ways of interpreting the claim such that it comes out false, but a charitable interpretation picks up on the fact that the utterer is talking, not about the persistence of persons, or of biological organisms, but the persistence of constituting atoms. Since these atoms were formed by fusion reactions in stars, then there's a sense in which it's true that our current temporal parts are mereological sums of stardust.

But even if you disagree with that particular controversial claim, my overall point is that, when focusing on the question of personal identity, we are interested in the persistence conditions for persons, not biological organisms or atoms. But what is a person?

This brings me to the second step in my argument: I hold that the concept of a person is a *moral* concept. First, a preliminary point: we all agree that persons have moral worth, but some non-persons have moral worth as well – it's wrong to torture a cow, even though a cow is not a person. But my main point is that persons, in addition to having moral worth, are (at least rudimentary) *moral agents*. Persons are morally responsible for the bad things they do, and morally commendable for the good things they do, and this is part of what is involved in being a person – it is to be the sort of agent to whom moral responsibility and commendability is to be attributed.

Without moral facts, it's unclear what does ground personal identity facts. Consider a full specification of what happens in some personal identity thought-experiment. We can specify where every particle goes, we can specify what the psychological traits are of every temporal part of persons involved, and we can specify how souls are persisting, if indeed souls exist. Given all these facts, which mereological sums of temporal parts constitute a single persisting person? One still has to know (for example) that the persistence of consciousness is what establishes the persistence of a person (as for example Locke thought), or that the persistence of the soul establishes the persistence of the person (Locke famously thought otherwise). But what grounds those facts?

Now, it could just be a matter of our individual or communal practices, or our arbitrary semantic conventions, that "person" refers to a certain type of entity. I'll consider (and reject) those lines of thought in the next section. But my position is that the fact that some temporal parts of entities have objective moral responsibility relations to other temporal parts enables us to pick out a certain privileged mereological sum of temporal parts, and that is the entity that should be thought of as a persisting person.

That, in a nutshell, is my argument. But before taking up the alternative lines of thought in the next section, there are four points I want to make.

First, I've said that I'm interested in the moral responsibility relation, and that this grounds personal identity. Specifically, what I'm interested in is the ancestral of the moral responsibility relation. Just as there are legal statues of limitations, so there may well be moral ones. Eighty-year-old Sarah is not morally responsible for three-year-old Sarah's gum-stealing, but that is not a problem, on my view. As long as temporal part #*n* of Sarah is morally responsible for something that temporal part #*n* – 1 did, and temporal part #*n* – 1 of Sarah is morally responsible for something that temporal part #*n* – 2 did, and so on, then all these temporal parts are parts of the same person.

Here's my second point. I take it that one can be morally responsible, not just for bad things one does, but for good things as well. The personal identity relation isn't just grounded in the morally wrong things people do. Even if a person never did anything morally wrong, the person is nevertheless morally commendable for good things she did in the immediate past, and that chain of moral commendability is enough to ground personal identity. One might ask: what if a person has a morally neutral day, such that she doesn't do anything morally bad or good – how can my view make sense of her continuance as a person throughout that day? I hold that persons never have morally neutral days; they are constantly doing morally good (or bad) things. The reason is that one is (at least minimally) morally commendable for *omitting* to do bad actions. So, a temporal part of a person that sits quietly has omitted from kicking anyone in the shin, and hence immediately subsequent temporal parts of the person are (at least minimally) morally commendable for that omission.

My third point is that there's an epistemological issue lurking in this whole discussion, which I want to acknowledge just to set aside. How is it that we can *know* whether a person persists through time? In other words, how can we know the answer to these moral questions?

I don’t have any special insight to offer regarding how to answer these questions. In many situations, the personal identity facts are just obvious to us: it’s obvious that the temporal part of Obama who got elected president in 2008 is part of the same person as the temporal part of Obama who got reelected in 2012. (And it’s also the case, once one thinks about it, that these obvious-seeming claims could be false, due to skeptical worries. Perhaps Obama was abducted by aliens in 2010, and was replaced by a qualitative duplicate.) Similarly, some moral facts are just obvious to us (barring philosophical worries about the status of moral truths, an issue I’ll come back to below). So in some cases, it’s obvious that a person persists, just as it’s obvious that a person is morally responsible for something she did.

But there are other cases where it’s not at all obvious whether a person persists, just as it’s not at all obvious that a current temporal part of a person is morally responsible for something that a previous temporal part of a person did. In these situations, there’s nothing more we can do than to use our standard methods: we do careful philosophical reasoning, we try to reflect in a rational, unbiased way regarding how things morally seem to us, and so on. My point is that there is an objective fact of the matter about whether a temporal part of a person is morally responsible for something that another temporal part did, and that’s enough to metaphysically establish whether those temporal parts are parts of the same person, even in situations where it’s epistemically hard for us to figure out which moral facts, and hence which personal identity facts, hold.

My fourth and final point addresses the question: to what extent is my position, that morality grounds personal identity, original? Well, I see foreshadowing in Locke, but my reading of Locke is controversial. (I have a lot to say about different ways of interpreting Locke on metaidentity, and how my theory of metaidentity relates, but that is best saved for another paper.) The only philosophical work I've found that comes somewhat close to my view is a paper by Eric Wiland (2000). Wiland argues that our theory of moral responsibility should influence our theory of personal identity; to that extent we are in agreement. But the details of his view, and his corresponding argument, are quite different. For example, he gives a partial characterization of the personal identity relation, by giving two necessary criteria for whether a person persists. One is that, for Y to be the same person as X, Y must be bodily continuous with X. I wouldn't want to build bodily continuity in as a necessary criterion (though it may end up being true, because of the objective moral facts being what they are). The other necessary criterion that he gives for Y to be the same person as X is that "Y is quasi-responsible for some action of X's" (2000, 84). Unfortunately, Wiland does not give a precise definition of quasi-responsibility, and his discussion introducing the concept leaves me nonplussed.

3. Conventionalism and the Concept of a Person

I maintain that facts about persistence of persons are grounded in objective moral facts. But a contrasting popular view in the metaidentity literature is *conventionalism*: facts about persistence of persons just hold as a result of decisions that are made by an individual or a community. Consider, for example, David Braddon-Mitchell and Caroline West's endorsement of conventionalism:

> To survive, on our understanding, is to preserve whatever property a person's

> (or perhaps community's) person-directed practices are organized around. (Braddon-Mitchell and West 2001, 61)

For conventionalists, whether a person persists depends on what person-directed practices are followed by a person (or the community the person is a part of). So whether, for example, you survive teletransportation depends on whether you (or your community) holds funerals for people who enter teletransportation machines, or treats such machines as a normal mode of transportation.

I believe that conventionalism holds for everyday created objects, like ships. The question of whether a ship persists is, indeed, purely grounded in our practices that reveal what it takes to be a persisting ship.

But are persons relevantly like ships? I maintain that conventionalism for persons is a highly implausible position, precisely because we do not take facts about our survival to depend on what practices we choose to adopt. Facts about our survival are, intuitively at least, objective facts. A virtue of my view is that it can accommodate that. Given that the concept of a person is a moral concept, and that there are objective moral truths, then I maintain that the most plausible metaphysical view is that there are objective facts of the matter about whether a person persists, and whether a person persists is determined by the objective moral facts.

To further develop my point that the concept of a person is a moral concept, let's bring in a distinction that Matti Eklund makes in his underappreciated 2004 paper on personal identity. Eklund distinguishes between the "moral question" and the "semantic question" of personal identity. The moral question is:

> what is the nature of the entities we should focus our prudential concerns and ascriptions of responsibility around? (Eklund 2004, 489)

The semantic question is:

> what is the nature of the entities that ‘person’ is true of? (Eklund 2004, 489)

Regarding the moral question, Eklund recognizes that, in principle, the entities we should focus our prudential concerns around could be different from the entities we should focus our ascriptions of responsibility around. I maintain that ascriptions of responsibility are the focus of personal identity, and these ascriptions of responsibility should track the objective moral truths. Our prudential concerns have no such objective focus. (For example, I could be more prudentially concerned about my friend than myself when I risk my life to save her, and in choosing to be prudentially concerned in that way, I could be doing so in such a way that I don’t violate any objective dictates of prudence.) Thus, the aspect of Eklund’s moral question that I’m most interested in is:

> what is the nature of the entities we should focus our ascriptions of responsibility around?

Since we would want our ascriptions of responsibility to track the objective moral truths about responsibility, this question has the desired focus, linking the concept of a person to the objective moral truths.

But what about the semantic question – how is it that we actually use the concept ‘person’? Eklund argues that it’s not clear whether the answers to the moral question and the semantic question are the same. But even if they aren’t the same, Eklund points out that the semantic question ends up looking rather unimportant. His paper ends with the following:

> Suppose we have answered the moral question: we have figured out what person-like entities (if any) we should structure our … ascriptions of responsibility around. … Still, we have not yet answered the semantic question: in order to do so we must take a close look at what our conception of persons is. However, how important is what is missing? … All we would not know is which entities happen to be picked out by a particular concept of ours. The significance of this question seems to pale in comparison to the others. (Eklund 2004, 507-8)

I agree with Eklund that, if the moral question potentially has a different answer than the semantic question, then the moral question is the important one. In fact, the semantic question looks so unimportant when divorced from the moral question that I maintain that it is a mistake to so divorce it. Of course, different people understand a concept like ‘person’ in different ways, and there is debate about what exactly the concept amounts to. But a way to help resolve this debate is to make clear that what’s implicit in our concept of a person are attributions of moral responsibility, and hence, the answer to the semantic question is the answer to the moral question.

That concludes my discussion of the semantic question and the moral question of personal identity. But there’s one more thing I want to say about Eklund. Eklund argues that our concept of a person is indeterminate, because “our conception of what persons are is not such as to decide what to say with respect to some of the problem cases discussed in the literature” (Eklund 2004, 490). I would argue instead that the apparent indeterminacy arises, not because we don’t have a full account of the concept of a person, but because we don’t have agreed-upon answers to some of the problematic moral responsibility questions that can be asked in the context of a personal identity problem case. The concept of a person tracks moral responsibility; it's because there is no agreement on whether, for example, new Sally emerging from the teletransporter is morally responsible for something that old Sally did that one might mistakenly think there is indeterminacy in the concept of a person.

I have one final point to make about the concept of a person, before moving on. There are various definitions of the concept of a person given in the literature (in the context of the abortion debate, for example); how does my moral-responsibility-based conception of a person relate to those definitions? Consider, for example, Michael Tooley's account:

> Something is a person if and only if it is a continuing subject of experiences and other mental states that can envisage a future for itself and that can have desires about its own future states. (Tooley 1979, 91)

Definitions like this are compatible with my account of persons; they provide an answer to the moral question "what is the nature of the entities we should focus our ascriptions of responsibility around?" Based on Tooley's definition, the answer would be:

> The entities we should focus our ascriptions of responsibility around are the continuing subjects of experiences and other mental states that can envisage futures for themselves and that can have desires about their own future states.

Given that there are objective moral truths, and that our ascriptions of responsibility should track the objective moral truths, there is an objective fact of the matter regarding whether this answer is the right one. If the objective moral ascriptions of responsibility apply to all and only the continuing subjects of experiences and other mental states that can envisage futures for themselves and that can have desires about their own future states, then Tooley's account would be a correct account of the concept of a person. Thus, an account of personhood like Tooley's is not incompatible with mine; we're just engaging in different levels of analysis.

4. Dropping the Assumptions

My argument above was based on three assumptions; let's now consider what happens if we drop the assumptions. I'll start with the first assumption, that there are objective moral truths. There are three standard moral anti-realist positions one could endorse, relativism, nihilism, and non-cognitivism; I'll talk about how my metaidentity position would fare on each one.

According to moral relativism, moral facts are relative to an individual, or community (or something along those lines). Given my position that the moral facts establish the personal identity facts, it would follow that personal identity facts are similarly relative to an individual or community – the conventionalist position discussed above. Those who are inclined toward moral relativism may well be satisfied with this corresponding conventionalism in personal identity situations. Metaidentity conventionalism is typically argued for without bringing in moral relativism; I've presented a new line of argument for metaidentity conventionalism that moral relativists could give.

According to moral nihilism, all positive moral claims are false (because they are attributing moral properties, and moral properties don't exist). If we take this view that there are no positive moral facts, and apply it to my position that morality grounds personal identity, then we get the view that there are no positive personal identity facts. The claim that Sally survives the teletransporter is false – and indeed, all claims that a person persists are false.

According to non-cognitivism, moral utterances don’t have truth-conditions – they are, for example, like expressions of approval or disapproval, or expressions of emotion. One way to apply that to personal identity is as follows. Suppose that Fred utters: “Sally survives the teletransporter”. This is Fred expressing his desire to treat new Sally as if she is the same person as old Sally. Fred could be an enlightened non-cognitivist, and endorse the connection between morality and personal identity, and hence Fred could realize that there is no fact of the matter regarding whether Sally survives. In uttering “Sally survives”, Fred is simply expressing his desire to treat new Sally the same sort of way he treated old Sally. (For example, if Fred was married to old Sally when she entered the teletransporter, he is expressing his desire to continue to treat new Sally as his partner.)

To sum up: because of the link I’m endorsing between morality and personal identity, the different metaethical positions correspond to different metaidentity positions. All these metaidentity positions are prima facie live options, though my preferred one is the one that is based on moral realism.

Let’s now turn to the second and third assumptions: that the perdurance theory of persistence is true, and that the doctrine of unrestricted mereological composition is true. Since these two are linked, let’s drop them together: suppose that the endurance theory of persistence is true (an object is wholly present at every time that it exists), and that collections of objects don’t always compose a whole.

The most radical way to give up unrestricted mereological composition is to be a mereological nihilist. In that situation, persons wouldn't persist at all – unless they are metaphysical simples (a view defended by for example David Barnett (2013)). Let's suppose that the correct replacement for unrestricted mereological composition is such that it does allow persons to persist. But what grounds the fact that some particular person, Sally, continues to exist?

My answer, recall, is that all collections of temporal parts exist, and one of them is picked out as morally significant, and is identified as the persisting person. But the endurance theorist has to say that it's a metaphysical fact of the matter that a particular thing, Sally, continues to exist. It's unclear to me what (if anything) grounds those metaphysical facts. If nothing grounds those metaphysical facts, how do we have epistemic access to them? (We arguably have epistemic access to the moral facts via our faculty of rational intuition – do we have access to the ungrounded metaphysical facts about endurance via similar means?) And if something does ground those metaphysical facts about endurance, what is the ground? The endurance theorist is welcome to take on board my answer, that the moral facts are the ground. But if the endurance theorist does not make that move, then I find it mysterious what (if anything) is the ground for these metaphysical facts about whether a person continues to be wholly present, and how we could have access to these metaphysical facts.

Some endurance theorists argue that there is no interesting debate about personal identity to be had once one assumes the truth of the perdurance theory and unrestricted mereological composition. Eric Olson, for example, holds that "if material objects are temporally extended, then there are no substantive metaphysical problems about our identity through time, but only semantic questions about how the *language* of personal identity works" (Olson 1997, 5). (Later in his book, Olson says that another assumption is needed to get his conclusion that personal identity questions become semantic questions; this assumption is something like the assumption of unrestricted mereological composition: "that every matter-filled region of spacetime contains an object" with the extension of that region (Olson 1997, 162).)

Let's consider Olson's claim that, given my perdurance and unrestricted mereological composition assumptions, personal identity questions are just semantic questions. Olson's thought is that, if all the mereological sums of temporal parts are real, then it's just a semantic question of which ones count as being the referent of our concept 'person'. But what Olson is missing is that the objective moral facts can be brought in to ground our ascriptions of which mereological sums count as persisting people. It's true that there's a residual semantic question, but that holds for all metaphysics. If we redefine "God" as "love", then I'm no longer an atheist, because I believe in love. But the concept of God is a metaphysically important concept, as is the concept of a person. Because the concept is metaphysically important, then picking out which entities count as persisting people is metaphysically important; it's not just a semantic issue. And my key point is that one can hold that the concept of a person is metaphysically important even if one endorses the perdurance theory and unrestricted mereological composition, by recognizing the connection between personhood and the objective moral truths. (If moral anti-realism is true, then while the connection between personhood and morality is still there, the concept of a person is presumably less metaphysically important, because morality itself doesn't have an objective metaphysical status.)

5. Four Objections

I'll close out this paper by considering four objections to my argument.

5.1 Objection #1: Fetuses are People Too

Some philosophers hold that a (human) fetus is a person. But fetuses aren't morally responsible (setting implausible claims about original sin aside). Hence, a fetus is a persisting person without having moral responsibility relations between its temporal parts, and hence, my view of persistence of persons is false.

It's true that my theory of metaidentity is incompatible with some other views about what counts as a person and who is morally responsible. But that is not problematic, because those other views are false. A fetus is not a person, nor are recently born infants. Providing an argument for this is beyond the scope of this paper, but briefly, my reason is as follows: a fetus is not cognitively sophisticated enough to be morally responsible for its actions, and hence is not a person. But I'd also be happy to follow for example Tooley's reasoning: a fetus is not a continuing subject of experiences and other mental states that can envisage a future for itself and that can have desires about its own future states, and hence a fetus is not a person. Assuming that Tooley's account of personhood is correct, then once an entity becomes a continuing subject of experiences and other mental states that can envision a future for itself and that can have desires about its own future states, it follows (by my lights) that this entity is morally responsible for its actions (in at least a rudimentary way), and hence it follows that it is a person.

Note that it does not follow from what I've said that abortion or infanticide is morally permissible. Just as we have a prima facie moral obligation not to kill cats, even though cats are not people, so we may have a prima facie moral obligation not to kill non-person humans.

Given that fetuses are not persons, and older children are, how does the transition happen? Is it an all-at-once shift, or are there degrees of personhood? My view is as follows: if the agent goes from being not morally responsible to morally responsible (even if morally responsible just to a small degree), then that's enough for the agent to go from being a non-person to a person. But if there is somehow proto-moral responsibility (where it's not the case that the agent is simply not morally responsible, but it's not the case that the agent is definitely morally responsible), then there would be intermediate degrees of personhood as well, corresponding to these proto-moral responsibility states.

5.2 Objection #2: Personhood is More Fundamental than Moral Responsibility

Some might argue that a person can persist through time, even if there are no moral responsibility relations that hold between the temporal parts of the person over that time interval. Consider a person, Ally, who is by herself, on a desert island, and will continue to be by herself for the rest of her life. This is a special island such that it (and the surrounding waters) contain no sentient creatures other than her. These objectors could maintain Ally can't do anything morally wrong or right, because there are no sentient creatures for her to do those morally wrong or right actions towards. Ally is morally neutral, and yet we all agree that she continues to persist as a person. These objectors would conclude that personhood is more fundamental than moral responsibility.

I'll present five possible ways of replying to this objection. First, theists could hold that Ally is always in a relationship with God, and hence can always do morally wrong or right by God. Second, one could hold that Ally can still harm herself, and that in itself is morally wrong. (Similarly, Ally can refrain from harming herself, and that is morally commendable.) Third, one could endorse something like virtue ethics, which holds that Ally's developing a good moral character is itself morally commendable, even if there is no one (other than Ally) to benefit from her having that good moral character. So, if a temporal part of Ally cultivates the desire to commit genocide, then the genocide-desiring immediately subsequent temporal parts of Ally are morally responsible for having cultivated that desire. Similarly, if a temporal part of Ally refrains from cultivating that desire, then the immediately subsequent temporal parts are morally commendable for having refrained. This establishes the appropriate chain of moral responsibility and commendability relations, and that grounds Ally's persistence as a person through time.

The second and third answers are my favorites, but if you don't like them, here are two more. One could appeal to counterfactuals: Ally persists because, had other people been around, she would have behaved in a morally accountable way toward them. Or, one could appeal to capacities: Ally persists because she has the capacity for being morally responsible. By my lights, those last two options bring in too much metaphysical baggage, but if you're not happy with the other answers, feel free to consider taking on the baggage.

5.3 Objection #3: What About Sleep? What About Comas? What About Brainwashing?

Some might argue that a person can't be morally responsible for anything while she is asleep, but she clearly does persist as a person while she is asleep. Thus, my grounding of personal identity in morality fails.

The objector here is forgetting that one can be morally responsible not just for actions, but also for omissions. Omissions are key for the sleep situation. Bob is currently asleep, but he could have woken up just before now and done a morally wrong act; Bob is morally commendable for omitting to wake up and do something morally wrong.

But what about comas? Can Bob persist as a person through a coma? I hold that, if Bob is in a coma, Bob is not morally commendable for omitting to wake up, since he is incapable of doing so. Thus, when Bob is in a coma, he continues to exist as a human organism, but not as a person.

Suppose that Bob's coma is temporary, and he eventually recovers from it. After the coma, Bob continues to exist as a person. Should we be bothered by Bob's discontinuous existence here? I will argue that we shouldn't, for two reasons.

First, we have some continuity, because Bob continues to exist as a human organism throughout the coma, even though Bob as a person goes out of existence and later comes back into existence. (The referent of "Bob" changes depending on context – sometimes it refers to the person, sometimes it refers to the human organism, sometimes it refers to stardust.)

Second, metaphysicians are already familiar with discontinuous existence, when for example a person enters a time machine in the year 2015 and instantaneously (from the standpoint of her personal time) and discontinuously appears in 1815. Of course, time travel is controversial, but the arguments that are used to justify the continued existence of a person over a discontinuous time travel jump are good arguments, and can carry over to justify the continued existence of Bob as a person. (Such arguments are given by for example David Lewis (1976a).)

For the final case where a person arguably continues to exist without being morally responsible, consider brainwashing. Suppose Sally is given a drug that gives her the desire to kill Fred – Sally is a person, but is arguably not morally responsible for killing Fred. Does this mean that Sally persists as a person without being morally responsible? No, because while Sally is brainwashed to kill Fred, Sally isn't brainwashed to (say) omit from kicking John. So when Sally does omit from kicking John, she's morally commendable for that action (even though she's brainwashed). It's only if the brainwashing controlled all of Sally's actions (and omissions) that she wouldn't be morally responsible – and if Sally is under that sort of complete control, then, indeed, she is no longer a person.

5.4 Objection #4: What About Fission?

So far, I've talked about personal identity thought-experiments as if they involve a temporal part of a person pre-experiment, and a temporal part of a person post-experiment, and we simply have to figure out whether those are temporal parts of the same person. But what about fission? Call "old Katie" the temporal part of a person that enters the fission machine, and "Katie1" and "Katie2" the temporal parts that emerge. How can my grounding of personal identity in terms of moral responsibility make sense of this scenario?

My answer is: unproblematically. Suppose first that the objective moral facts are such that Katie1 and Katie2 are *not* morally responsible for anything old Katie did. It would follow, based on my metaidentity theory, that the person which included the old Katie temporal part has ceased to exist, and two new people have come into existence, with qualitative similarities to the person who ceased to exist.

Now suppose that the objective moral facts are such that Katie1 and Katie2 *are* morally responsible for something that old Katie did. It follows, by my analysis, that Katie1 and Katie2 are parts of the same person as old Katie. How can one make sense of this? On the perdurance theory of persistence, at least, this is unproblematic, as Lewis (1976b) has shown. There is a persisting person that the temporal parts old Katie and Katie1 are a part of, and there is a different persisting person that the temporal parts old Katie and Katie2 are a part of. The four-dimensional spacetime worms of these persisting people overlap, sharing some temporal parts in common.

But what if (as many believe) proponents of the endurance theory cannot make sense of a person persisting through fission? If the objective moral facts are such that Katie1 and Katie2 are both morally responsible for something that old Katie did, then by my lights it follows that Katie did persist through fission, and hence the endurance theory is false. In this way, we can look to morality not simply to provide answers to questions about personal identity, but to provide answers to other perennial questions in metaphysics too.[1)]

[1)] For helpful comments, thanks to anonymous reviewers for this journal, and audiences at University of Sydney, the 2012 Australasian Association of Philosophy conference in Wellington, and University of Colorado Boulder. Also thanks to Rebecca Chan, Matti Eklund, Nicole Hassoun, Ryan Jenkins, Brian Kierland, David Kovacs, Kristie Miller, Chad Mohler, Denis Robinson, and David Shoemaker.

References

Barnett, David (2013) "You Are Simple", in Georg Gasser and Matthias Stefan (eds.), *Personal Identity: Complex or Simple?*, Cambridge: Cambridge University Press.

Braddon-Mitchell, David, and Caroline West (2001) "Temporal Phase Pluralism", *Philosophy and Phenomenological Research* 62, pp. 59-83.

Eklund, Matti (2004) "Personal Identity, Concerns, and Indeterminacy", *The Monist* 87, pp. 489-511.

Huemer, Michael (2005) *Ethical Intuitionism*, New York: Palgrave MacMillan.

Lewis, David (1976a) "The Paradoxes of Time Travel", *American Philosophical Quarterly* 13, pp. 145-152.

Lewis, David (1976b) "Survival and Identity", in Rorty, Amelie O. (ed.), *The Identities of Persons*, Berkeley: University of California Press, pp. 17-40.

Locke, John (1694) *An Essay Concerning Human Understanding*, Second Edition.

Olson, Eric (1997) *The Human Animal: Personal Identity Without Psychology*, Oxford: Oxford University Press.

Rosen, Gideon (2010) "Metaphysical Dependence: Grounding and Reduction", in Bob Hale and Aviv Hoffmann (eds.), *Modality: Metaphysics, Logic, and Epistemology*, Oxford: Oxford University Press, pp. 109-135.

Shoemaker, David (1999) "Utilitarianism and Personal Identity", *Journal of Value Inquiry* 33, pp. 183-199.

Tooley, Michael (1979) "Decisions to Terminate Life and the Concept of a Person", in John Ladd (ed.), *Ethical Issues Related to Life and Death*, Oxford: Oxford University Press, pp. 62-92.

Wielenberg, Erik (2009) "In Defense of Non-Natural, Non-Theistic Moral Realism", *Faith and Philosophy* 26, pp. 23-41.

Wiland, Eric (2000) "Personal Identity and Quasi-Responsibility", in T. van den Beld (ed.), *Moral Responsibility and Ontology*, Dordrecht: Kluwer Academic, pp. 77-87.

Philosophy Department, Univ. of Colorado at Boulder,
bradley.monton@colorado.edu

철학적 분석 31 (2014) pp.27-47

The Truth Requirement on Evidence

Douglas Roland Campbell

In this essay, I argue that one's intuitions about evidence require that only what is true be counted as evidence. I present this prima facie case for a truth requirement on evidence in a few steps. First, I begin by outlining exactly what is, intuitively, expected of evidence. Then, I offer an exposition of two competing views. The first respects a truth requirement; the second does not. After this exposition, I argue that the second is beset by conflicts with one's intuitions. I additionally argue that the second cannot perform a large majority of the tasks intuitively demanded of evidence that I outlined earlier in the essay. The first view is not met with any of these problems. The results generalize: accounts that omit a truth requirement oppose one's intuitions. Lastly, I present a case in which there, intuitively, seems to be false evidence, and then I argue that it does not, in truth, pose a threat to this view.

Keywords Evidence; justification; epistemology; truth; Williamson

One’s intuitions about evidence make a clear demand: evidence is required to be true. This request is made by those intuitions which disclose the functions of evidence. These intuitions insist that pieces of evidence play a role in explanations, that they determine probabilities, that they help form inferences, and that they assist in satisfying the norm that one ought to regulate one’s degree of belief in some proposition in accordance with one’s evidence for that proposition. The possibility of false evidence threatens these functions, all of which are important to doxastic life. Moreover, a conception of evidence without a truth requirement has two unintuitive consequences: the first is that evidence is able to exclude truths, and the second is that inconsistent bodies of evidence are allowed. In sum, without a truth requirement, a picture of evidence does not just sit uncomfortably with one’s intuitions, but it actively opposes them. For these reasons, one’s intuitions prohibit false evidence.

I. The Functions of Evidence

Doxastic agents use evidence to accomplish a variety of assignments in their lives. Thomas Kelly writes that “intuitively, one’s evidence is what one has to go on in arriving at a view” (Kelly 2008, 942). This broad portrait of evidence’s responsibilities can be parsed into a handful of specific roles. Firstly, evidence is the sort of thing one uses to ascertain the probability of a given hypothesis being true. Evidence is used to confirm and disconfirm theories to different degrees. Evidence can outright verify a theory by “conclusively [establishing] it;” other times, it can outright falsify a theory by conclusively establishing “that the theory in question is false” (Kelly 2008, 933). These two extremes are continuous with a wide spectrum of degrees of confirmation in between. One similar function, which reflects the possibility of falsification, is that evidence can altogether eliminate and rule out hypotheses.

Another important and related fact is that evidence features heavily in explanatory narratives. This feature is crucial in providing inferences to the best explanation, which is a technique used by one who "seeks to formulate hypotheses that provide good explanations for one's evidence" (Brueckner 2005, 436). When a hypothesis compellingly *explains* some instance of evidence, the hypothesis gains clear support. One indirect way a theory may receive such support is when a given piece of evidence disconfirms or, better still, falsifies "some otherwise formidable rival theory" (Kelly 2008, 934). Evidence has a place in explanations and, in this way, continues to be invaluable in one's doxastic career.

Evidence's behavior in this regard is conducive to another of its functions: it can determine to what extent some proposition μ warrants one's assent, by evaluating how likely it is that μ is true, on the strength of the support it derives from one's evidence. The ability of evidence to play this part accommodates a normative aspect of doxastic life: one ought to regulate one's beliefs in accordance with one's evidence. Two things are worth noting here. What is reasonable for one to believe will be a "highly relativized matter," since the collection of evidence owned by one doxastic agent will not always be identical to a collection of evidence owned by another (Kelly 2008, 939). Moreover, as one's body of evidence changes, so does what ought to be considered reasonable. Specifically, Kelly maintains that "what it is reasonable for one to believe on the basis of [one's] evidence undergoes" changes corresponding to increases in one's inventory of evidence (Kelly 2008, 937). In light of the possibility of *change* in bodies of evidence, it is intuitively important that any conception of evidence be such that one can *update* probabilities in accordance with shifts in one's total body of evidence. Accordingly, the norm that one ought to regulate one's beliefs in accordance with one's evidence can be satisfied by doxastic agents, with reference to their own evidence as it changes over time. Intuitively, this job of evidence is just one item on evidence's list of tasks, amongst others, such as determining probabilities, being a constitutive ingredient of explanations, and, lastly, forming inferences.

Evidence plays a familiar role in inferential operations. Here, as in explanatory moments, evidence is part of a narrative. To illustrate, consider the case of Sherlock Holmes. After he gathers evidence from, say, a crime scene, Holmes "infers the identity of the person who committed the crime" (Kelly 2008, 942). Disclosed in this intuition about how evidence is used by Sherlock Holmes is the plain datum that evidence sometimes functions as premises in inferences (and, so, is propositional – at least some of the time). Bundled together, the roles of evidence presented thus far amount to what was stated earlier: evidence is what one has to go on in arriving at a view. Perhaps another way of understanding this job description is that evidence assists one in reaching a *conclusion*. This approach has, if nothing else, the advantage that it makes clear the task accomplished by evidence in the formation of inferences.

One special quality of these assignments in doxastic life, carried out by evidence, is that they cannot be so carried out if evidence is not always true. The only exception is the way evidence features in inferences, which is feasible with or without a truth requirement. I shall argue that, since the functions of evidence enumerated here are given in intuitions, these same intuitions militate in favor of a truth requirement on evidence. One's intuitions coalesce and arrange a *prima facie* case for such a truth requirement. I shall proceed by presenting two competing conceptions of evidence: the first respects a truth requirement, and the second does not. The absence of a truth requirement will pose special problems for the second conception, which, I allege, — since the problems stem from nothing other than the permissibility of false evidence — generalize: all theories of evidence without a truth requirement are engaged in a deep conflict with one's intuitions.

II. The Exposition of *E* = *K*

The first view is one on which there exists a truth requirement on evidence: there is no such thing as false evidence. On this account, the set of one's evidence is extensionally equivalent to the set of one's knowledge.[1] The proponent of this view — *E* = *K* — supposes "that knowledge, and only knowledge, justifies belief" (Williamson 2000, 185). Subject *S*' knowledge can be used as evidence for different hypotheses under consideration by *S*. Specifically, *e* is evidence for hypothesis *h* for *S* if and only if *S*' evidence includes e — so, *S* knows *e* — and the probability of *h*, conditional on *e*, is higher than the prior probability of *h* (Williamson 2000, 187). Under *E* = *K*, one's evidence includes all of one's knowledge, and all of one's knowledge is available as one's evidence.[2]

[1] Formally, $\forall x((x \in K) \longleftrightarrow (x \in E))$.

[2] Notably, all evidence here is propositional. In fact, this stance regarding evidence forms some of the background of my entire argument. It is by no means uncontroversial, but it is an issue that must simply be dealt with elsewhere. Briefly, I shall make recourse to Williamson and say that this understanding of

Only knowledge suffices as evidence. The essence of $E = K$ is that, for example, “if I do not know that the mountain is that shape, then that it is that shape is not part of my evidence” (Williamson 2000, 200). The thought that one could stand in a relation to one’s evidence other than a knowledge-relation is pernicious. If *e* were an element of the set of subject *S*’ evidence by virtue of *e*’s “good cognitive status short of knowledge, then a critical mass of evidence could set off a chain reaction” (Williamson 2000, 201). An implausible amount of propositions would qualify as evidence in such a case.[3] Additionally, the restrictions put on what is evidence — such that knowing *e* is a sufficient and necessary condition for *e* being evidence — are motivated by the functions that evidence performs in one’s doxastic life. I will argue that, on $E = K$, one can rule out hypotheses that are incompatible with one’s evidence, because one’s evidence is *known* and, by the factivity of knowledge, true. Moreover, one can reason with Bayesian probabilities to ascertain just how much likelier a certain hypothesis is made by the truth of one’s evidence. The consistency of one’s body of evidence guaranteed by $E = K$ is a necessary condition for the possibility of Bayesian

evidence is not academic or unintuitive, for “even in the courts, the bloodied knife provides evidence because the prosecution and defence offer competing hypotheses as to why it was bloodied or how it came into the accused’s possession; the evidential proposition is *that* it was bloodied or *that* it came into the accused’s possession” (Williamson 2000, 195). Propositions are just the sort of things that one uses as evidence. Objects may be the source of a diverse number of propositions, which, in turn, are evidence, but they themselves are not evidence.

[3] Williamson’s argument here is instructive: placement of a truth requirement on evidence does not mean that all one’s intuitions are satisfied. Indeed, a world in which one’s justified true beliefs counted as evidence would be a world in which there is an intuitively implausible amount of evidence, yet a truth requirement would still be observed. The takeaway is that a truth requirement is a necessary, but insufficient, condition for satisfying one’s intuitions regarding evidence.

reasoning. Lastly, one's evidence allows one to regulate one's beliefs in accordance with their evidential support. It takes a truth requirement on evidence to vindicate these doxastic practices.

III. The Exposition of *E* = *NPJ*

On the other hand, Alvin Goldman's account ─ *E* = *NPJ* ─ does not restrict one's evidence to truths only. On *E* = *NPJ*, the set of one's evidence is co-extensive with the set of propositions one is "non-inferentially, propositionally justified" in believing, even if one does not believe them (Goldman 2009, 86).[4] Goldman's defense of *E* = *NPJ* consists in efforts to demonstrate that *E* = *K* can be matched. He endeavors to show that there exists at least one opponent to *E* = *K* that can equally meet the intuitive demands that are put on evidence (Goldman 2009, 85-86). It would follow that an account without a truth requirement could satisfy one's intuitions just as much as conceptions which outlaw false evidence. Goldman argues that, on *E* = *NPJ*, evidence can be used to rule out hypotheses and that it can be used to regulate the degree to which one believes some proposition, in proportion to the evidential support for the proposition. If he is correct, then he has presented a coherent and equally appealing alternative to *E* = *K*.

It is the vacuum of a truth requirement on evidence that makes E = NPJ problematic. If all the evidence in question is true, then *perhaps E* = *NPJ* does, in fact, do everything Goldman enumerates.[5] The challenge

[4] Goldman explains that by 'propositional justification', he means "the sort of justification an agent has vis-à-vis a proposition when she is justified *in* believing it, whether or not she actually believes it" (Goldman 2009, 86).

is that not all evidence, on *E* = *NPJ*, is true. As Timothy Williamson states, there is a primordial difference between *E* = *K* and *E* = *NPJ*, which is that “on E = K all evidence is true, since knowledge is factive, while presumably on *E* = *NPJ* some evidence is false” (Williamson 2009, 308). I will argue that, on *E* = *NPJ*, evidence excludes truths, permits inconsistent bodies of evidence, no longer allows for updating beliefs in proportion to evidence vis-à-vis Bayesian probabilities, and, broadly, does not always play the roles that one needs evidence to play. Only true evidence can meet those demands.

IV. To the extent that E = NPJ does not observe a truth requirement on evidence, unintuitive facts are true of one’s set of evidence

E = *NPJ*, not observing a truth requirement on evidence, allows two unbearably unintuitive facts about evidence: first, evidence excludes truths, and, second, agents can collect inconsistent bodies of evidence. To demonstrate the possibility of the first, let us suppose that subject *S* is non-inferentially, propositionally justified in believing ‘there is a computer screen before me’; let us use ‘*ϕ*’ to name this belief. On Goldman’s view, this belief is a member of the set of *S*’ evidence (Goldman 2009, 86). One can imagine a possible world in which *ϕ* is non-inferentially, propositionally justified for *S*, yet *ϕ* is false. Still, for Goldman, *ϕ* would be part of *S*’ evidence. If *S* were to believe falsely

[5] If *E* = *NPJ* is defective in a way not entailed by the possibility of false evidence, then I shall be silent on it. The discussion of *E* = *NPJ* is included only insofar as it facilitates a larger discussion of how the omission of a truth requirement brings a conception of evidence out of line with one’s intuitions.

‘there is a computer screen before me’, this belief could be used to eliminate altogether the true proposition ‘there is no computer screen before me’.[6] Accounts of evidence that omit a truth requirement allow evidence to exclude truths.[7] This unintuitive property of evidence is true not just of *E* = *NPJ* but of any account that omits a truth requirement. *E* = *NPJ*, as one such account, is illustrative of a general truth.

Goldman disagrees that allowing falsehoods as evidence rules out certain truths. In his view, the endowment of falsehoods with the standing of evidence “does not entail that some truths are permanently consigned to epistemic oblivion” (Goldman 2009, 88). He supports this assertion by arguing that a belief can have its membership in *S*’ set of evidence revoked at a later time; changing sets of evidence will allow for some truths to eventually be *recovered*. Consequently, one’s evidence may exclude a truth, but not forever. This proposal may, at first, seem enticing. Earlier, it was pointed out that one’s catalogue of evidence can change over time. Goldman’s proposal seems to reflect this integral fact about evidence. However, his qualification that truths are not *permanently* excluded does not capture what was unintuitive about *E* = *NPJ*, and, therefore, it fails to resolve the difficulty. What is implausible is that one’s evidence should exclude a truth *ever* ― even if that exclusion is just temporary. Additionally, just because the false belief’s status as evidence is temporary does not mean it *will* be revoked. A particular truth

[6] Goldman does maintain that the set of one’s evidence includes those propositions which one is non-inferentially, propositionally justified in believing, *even if one does not believe them* (Goldman 2009, 86). I have just designed this specific case as one in which the (purportedly evidential) proposition `there is a computer screen before me’ is believed.

[7] My use of ‘exclude’ will be identical to my use of ‘rule out’ and to my use of ‘eliminate’.

may, for a certain agent, be ruled out for his or her entire life, even though the good evidential standing of the false belief that excluded it was, in some sense, revocable. It is precisely this difficulty that Goldman's suggestion fails to resolve.

On the other hand, one could dissolve these challenges by imposing a truth requirement on evidence. Williamson's proposal — $E = K$ — accomplishes this feat. Given the factivity of knowledge, evidence never clashes with a truth. One's evidence is always true. Even so, there are instances in which a subject's "evidence may make some truths improbable, but it should not exclude any outright" (Williamson 2000, 201). This feature of one's evidence removes the difficulty present in Goldman's rival account. A piece of evidence *e*, on Williamson's view, may lower the probability of *h*; the prior probability of *h* may be higher than the probability of *h*, conditional on *e*, but *h* is never ruled out by one's evidence. An example may be the case in which Julia kills Chris and plants misleading clues in an attempt to frame Robert. An investigator may find, say, the murder weapon in Robert's room; accordingly, one member of the investigator's set of evidence will be 'the murder weapon was found in Robert's room', which can be used to increase the probability of the false hypothesis that Robert did it. It would even lower the probability of, but not exclude altogether, the truth that Julia killed Chris. Evidence can make the hypothesis *h* less likely, but it would never rule *h* out. Consequently, with a truth requirement on evidence, one avoids the exclusion of truths.

Nevertheless, there are times when one treats some false belief *υ* as though it were evidence. I could falsely believe that *υ* is part of my evidence. This false belief *υ* could lead me to exclude some truths. In reality, though, "no true proposition is inconsistent with my evidence, although I may think that it is" (Williamson 2000, 201). The belief '*e* is part of S' evidence' is not a sufficient condition for *e*'s membership in *S*' set of evidence.[8] It does not follow from a belief that *e* is part of one's evidence that *e* is part of one's evidence (Williamson 2000, 191). One may treat a false belief *υ* as part of one's evidence, but, in light of a truth requirement, no problem emerges: *υ* is simply not evidence. One can have mistaken beliefs about one's evidence, on *E* = *K*.

Another unintuitive feature of *E* = *NPJ* is that the omission of a truth requirement on evidence permits a stock of evidence that is not just at odds with some truths, but also with itself. It is inconsistent, which is unintuitive in its own right and, furthermore, generates obstacles when calculating probabilities. The determination of probabilities is especially consequential, since one's evidence *e* supports hypothesis *h* only if the probability of *h*, conditional on *e*, is greater than the prior probability of *h*. Williamson presents a case in which *S* is examining a perceptual illusion: a "twisted closed loops of stairs" (Williamson 2009, 310). Here, *S* is non-inferentially, propositionally justified in believing that T_1 is under T_2 and, simultaneously, that T_2 is under T_1. Since both propositions satisfy the necessary and sufficient conditions for membership in *S*' set of evidence, under *E* = *NPJ*, *S*' set of evidence can be inconsistent.

Even if this feature of *E* = *NPJ* were not unintuitive by itself, it would nevertheless frustrate efforts to discern which pieces of evidence supported which beliefs. Williamson insists that "there are grave difficulties in making sense of evidential probabilities on inconsistent

[8] Though not entirely pertinent, it is, for its own sake, worth pointing out that *S* holding the belief '*e* is part of *S*' evidence' is not just an insufficient condition for *e* belonging to *S*' evidence, but it is an unnecessary condition as well. One could, in other views not discussed here, have *e* in one's evidence without knowing, or even merely believing, that one does.

evidence, since conditional probabilities are usually taken to be undefined when conditioned on something inconsistent" (Williamson 2009, 310). Proposition *p* has a probability of 1 when it is conditional on itself, and a contradiction has a probability of 0 when it is conditional on anything at all. These two properties do not hold for a set of inconsistent pieces of evidence: if *p* were a contradiction, it could not be assigned a probability of 1, even when conditional on itself — because a contradiction has a probability of 0 when it is conditional on anything. However, with that fact in mind, one could assign *p* a probability of 0, but then it would not have a probability of 1 — even when conditional on itself. A necessary condition for *e* being evidence for, specifically, *h* is that it increases the probability of *h*. On Goldman's account, the calculation of probabilities is complicated. These complications can be obviated by attaching a truth requirement to evidence, which also precludes inconsistent evidence, for no truth is inconsistent with other truths.

V. To the extent that *E = NPJ* does not observe a truth requirement on evidence, it no longer performs the tasks that one, intuitively, asks evidence to perform

With untrue evidence, *E = NPJ* no longer functions as a competitive rival to *E = K*, just because it fails to satisfy one's intuitions about what evidence does. The first practice that allegedly both *E = K* and *E = NPJ* could do was allowing a subject to prefer *h* over competitor h_*. Goldman's justification of this activity, while adhering to *E = NPJ*, falls apart without a truth requirement on evidence. He argues that when S is propositionally justified in believing *e*, *S* will then, typically, believe *e* is true (Goldman 2009, 86).[9] In this case, the presence of propositional justification for *e* should give *S* good reason to prefer *h* over h_*, if *h* explains e better than h_* does. This line of reasoning is plausible. However, Goldman would have to clarify under what circumstances *h* explains *e* better than h_* does — or what it means to offer an explanation at all.

An appeal to Bayesian probability may have sufficed, but, as I argued earlier, certain problems hinder probability-calculations, given the permissibility of inconsistent bodies of evidence. Moreover, Goldman holds that "if *e* is true, and if *h* explains its truth better than h_* explains it, then *h*'s capacity to explain *e* confirms *h* for us. There is no need to postulate *E* = *K* to explain this phenomenon" (Goldman 2009, 86). There is something suspect about the antecedent of the conditional: the truth of *e* ought to be immaterial for Goldman. What he is confessing here is that truth does help, in some capacity, to make hypothesis-preferring possible and intelligible. This commitment is unusual, in light of the lack of a necessary connection, for Goldman, between evidence and truth.

Let me anticipate a reply from Goldman. The content of his objection to this criticism may be that he has nowhere written that the truth of *e* is a necessary condition for hypothesis-preferring. Rather, he just listed it as part of a sufficient condition. Even still, no matter what the nature is of truth's place in *E* = *NPJ*, truth is not secured by *e*'s status as evidence. It, therefore, is not obvious why an advocate of *E* = *NPJ* would have included a truth condition as part of the antecedent of the conditional

[9] Goldman's reasoning here is quite subtle: he suggests that when, say, Jessica is propositionally justified in believing *φ*, she will, "commonly (though not inevitably)," believe that she is propositionally justified; using this belief that she is justified, she will believe *φ* (Goldman 2009, 86).

cited above. However, if Goldman were to continue to insist that *e* being true is part of a merely sufficient condition and is not necessary for hypothesis-preferring, he still would have to hold forth on what it means to *explain better* without Bayesian probability. There is no such problem for the supporter of a truth requirement on evidence.

The second practice rendered impossible by the absence of a truth requirement is belief-regulation. Williamson writes that it is "hard to see why the probability of *h* on *e* should regulate our degree of belief in *h* unless we know *e*" (Williamson 2000, 200). He motivates this claim by appealing to the fact that when *h* and *e* are incompatible, one can rule out *h* if and only if *e* is known (Williamson 2000, 200). Goldman believes that *e* being known is not a necessary condition for ruling out *h* when the two are incompatible, and that *E* = *NPJ* can contend with *E* = *K*. In the end, though, Goldman's arguments are only plausible when one affixes a truth requirement to evidence. He affirms that one can regulate one's "degree of belief in *h* by the probability of *h* on *e*" (Goldman 2009, 86-87). Again, this recourse to probability is suspect: unless Goldman articulates how to salvage Bayesian probability when inconsistent evidence is permissible, these explications fail. Only now can one see that the content of the last section prefigured the content here; the unintuitive properties of evidence, according to *E* = *NPJ*, that do not directly hinder its ability to complete its many assignments *can* be brought to bear here. The permissibility of inconsistent bodies of evidence may have been, intuitively, a threat to *E* = *NPJ* in its own right, but the complications that frustrate Bayesian calculations are obstacles in the way of such (pretheoretically) plausible doxastic practices as belief-regulation and hypothesis-preferring. Predictably, one could dissolve these difficulties just by appending a truth requirement to evidence.

There may have been a sense that this practice — belief-regulation — is not always possible even under an account that observes a truth requirement. One does not always know whether a proposition is in one's evidence, just because sometimes one is "liable to misidentify the evidence for" some hypothesis (Williamson 1997, 722). Though the regulation of degrees of belief in accordance with how much support some belief receives from one's evidence — which can be determined by applying Bayes' theorem — is a normative component of doxastic life, "there is no infallible recipe for deciding in practice whether we know a proposition *p*," or, equivalently, on $E = K$, whether *p* is part of one's evidence (Williamson 2000, 191). The possibility of mistaken beliefs about one's evidence does not threaten this norm. Williamson proffers an analogy between the rules 'proportion your belief in *p* to your evidence for *p*' and 'proportion your voice to the size of the room' (Williamson 2000, 192). One may not always know what the size of the room is, but it does not follow that the rule does not hold. The rule in question is not descriptive, but, as a rule, it is a norm. Intuitively, one ought to regulate one's degree of belief in proportion to one's evidence; this rule is distinct from the claim that one, in fact, *does* regulate one's degree of belief in proportion to one's evidence. One may often make mistakes when attempting to follow the rule, but those errors are not equivalent to following "a different rule" (Williamson 2000, 192). Therefore, $E = K$ does not exclude anything intuitive about evidence.

Furthermore, in the absence of a truth requirement, *e* cannot eliminate the possibility of *h*. Goldman answers that "*e* does not have to be known (with all the baggage knowledge entails) for an incompatibility between *h* and *e* to rule out *h*. It suffices for *e* to be true" (Goldman 2009, 87). The trouble with Goldman's argument here is that the truth of *e* is not guaranteed by its status as evidence, under *E* = *NPJ*. Again, Goldman may maintain that this fact is no true trouble for him, because although it suffices for *e* to be true, it is not necessary for *e* to be true (and, so, he does not have to admit there is a truth requirement on evidence). There is, however, good reason for believing that *e* does have to be true to rule out hypotheses.[10] If one could rule out, using *e*, *h* even when *e* is false, then one would be able to rule out *h* when *h* is true. The exclusion of truths is one of the unintuitive features I, earlier, argued was true of evidence, under *E* = *NPJ*. As long as Goldman does not believe that truth is necessary to rule out hypotheses, he must admit, against intuitions, that evidence can rule out truths. If he, in reaction, holds that only truths can rule out hypotheses, then he is admitting that it is necessary for evidence to be true to perform this job. He would be placing a truth requirement on evidence. Plainly, then, *e* can rule out *h* only if *e* is true; if not, then evidence can exclude truths. A truth requirement is necessary for any account of evidence to cohere with one's intuitions about evidence.

VI. Presentation of an alleged counter-example and a reply

One potential counter-example to the intuitive nature of the truth requirement is a case in which a doxastic agent appears to reach, from false premises, a true conclusion, and, simultaneously, one intuits that the belief in the true conclusion is in good standing. It seems, after all, that

[10] I shall not elaborate on whether it is also sufficient for *e* to be true to rule out *h*. It may, for instance, be sufficient and necessary for *e* to be known to rule out *h*; Williamson holds such a position. However, my purpose here is only to write about a truth requirement, that is, a necessary condition.

the false premises have functioned as evidence. One would have to either throw out the good standing of the belief in the inferred proposition − which, again, is supported by one's intuitions − or admit the possibility of false evidence. So construed, these cases present a counter-example to my thesis that a truth requirement on evidence is demanded by one's intuitions. As such, it is incumbent upon the proponent of a truth requirement on evidence to show that what appears to be happening in this case is nothing more than an appearance.

I shall briefly outline an instance of this type of case. A sick man named Martin visits his doctor, Samantha, who instructs him to get eight hours of sleep each night. Always one to take the doctor's counsel seriously, Martin goes to sleep at eleven o'clock. However, he wakes up at one point and looks at his clock, which reads "3:30 am." Martin reasons from the premise that he has been asleep a mere four hours and thirty minutes to the conclusion that he has not yet satisfied Samantha's advice and gotten the prescribed eight hours of sleep. Unfortunately, Martin's premise is false: it is the appropriate time of the year to change one's clocks. His clock changed properly, automatically, during the night, and, so, he has, in fact, been asleep five hours and thirty minutes.[11] He used what supposedly constitutes false evidence to reach the true conclusion that he had not yet followed Samantha's advice. Accordingly, here is meant to be a case of false evidence, presented to disabuse a person of the impression that truth is a requirement for being evidence.

However, there are some good reasons to think that that is not what is happening. The force of this counter-example is felt only under a few

[11] This case is one, supposedly, of knowledge from falsehood. I have changed only small and negligible details from a nearly identical case, which I attribute here to Ted A. Warfield. cf. (Warfield 2005, 407).

conditions. Firstly, the proponent of a truth requirement must admit that the belief in the conclusion — in this case, Martin's belief that he has not yet slept the full eight hours — is in good standing, at least intuitively. Since I am arguing that one's intuitions demand that evidence always be true, a scenario in which there is an intuitively good belief formed using false evidence would be a powerful counter-example. For the sake of argument, I shall not contest that one's pretheoretical judgments point to the protagonist's belief, reached on the basis of false premises, being in good standing.

Secondly, to feel the force of the counter-example, one would also have to admit that Martin's premises were, in fact, satisfying a role played by evidence. One could simply deny that what Martin was doing — making an inference — is at all something which involves evidence. However, this strategy seems unattractive, because it was already mentioned earlier that one intuitive job of evidence is to function as premises in inferences, exactly in the same way Martin forms his inference in this case. As such, the application of this second strategy — in order to resist the thrust of the counter-example that one's intuitions do not always require evidence to be true — would force one to deny an intuitively appealing feature of evidence: namely, one uses evidence as premises. For this reason, it behooves me to opt not to use this second strategy.

The way forward, out of this counter-example, is to admit that Martin is treating his false belief as evidence, but that the treatment of a belief as evidence is not sufficient for it being evidence. In other words, "the subject [Martin] acquires knowledge because treating something as if it is evidence is a safe way of forming beliefs" (Littlejohn 2013, 159). This solution appeals to the proponent of a truth requirement for a few reasons. Firstly, it is consonant with the intuition that Martin's belief in the conclusion, in the above case, enjoys good standing. Secondly, it preserves the intuition that functioning as premises in inferences is something evidence does. Thirdly, and importantly, it trades on a distinction made earlier, which is also essential to the view being discussed here: treating a belief as evidence is neither identical to nor sufficient for the belief being evidence. There are many times when one treats some false belief as evidence. This treatment is not the same as it being evidence. It is essential for the proponent of a truth requirement to endorse this distinction. With this distinction in mind, one admits that Martin *knows* the conclusion of his inference, because, in some cases, treating a false belief as evidence produces knowledge.[12] Not *only* evidence can generate knowledge. Since treating a false belief as evidence is not a sufficient condition for it being evidence, it does not follow that Martin's false premise was false evidence. One's intuitions demand that there be no such thing.

VII. Conclusion

One's intuitions form a strong line of defense against the possibility of

[12] It is being assumed for the sake of argument that there is a way to reason from some false beliefs to knowledge. The defender of the truth requirement on evidence could have denied this assumption. It is important that the account of *how* one starts an inference with false beliefs and ends up with knowledge will be the same account of how false beliefs can function as evidence in a way that yields knowledge (yet denies that they are, in fact, evidence). Since the provision of this account is a burden that must be met by the believer in knowledge from falsehood (because, presumably, he or she needs to explain how this movement takes place), I do not need to specify how one can treat false beliefs as evidence and finish with knowledge. All I needed to do was accommodate supposed cases of knowledge from falsehood.

false evidence. So long as one respects the intuitions I have delineated, there is a strong case for a truth requirement on evidence. This conclusion is reached, just by virtue of the fact that the problems seen afflicting *E* = *NPJ* do so only because of its omission of a truth requirement. Without a truth requirement, evidence disappoints one's intuitions on multiple counts: where there is false evidence, there is the exclusion of truths, there is inconsistency in bodies of evidence ― with the attendant complications imposed on Bayesian calculations ― and there is just the general failure of false evidence to rule out and prefer hypotheses and allow subjects to regulate degrees of belief in accordance with the subjects' evidence. Only true propositions do what evidence needs to do. In other words, truth is a requirement on evidence. Furthermore, the distinction between treating a belief as evidence and a belief's status as evidence is made central, because it is on that basis that one can accommodate otherwise powerful counter-examples. Observance of a truth requirement is a necessary condition for being an account of evidence that is compatible with one's intuitions about evidence. It is in just this sense that one's intuitions demand evidence be true.[13)]

[13)] I would like to thank Julia Atack, Jessica Camps-Cooke, Gabriel Lariviere, Christopher Nguyen, Dirk Schlimm, and Solomon Tse for their helpful comments and feedback.

References

Brueckner, Anthony. (2005) "Knowledge, Evidence and Skepticism According to Williamson", *Philosophy and Phenomenological Research* 70, 2, pp. 436-443. Review of *Knowledge and Its Limits*, by Timothy Williamson.

Goldman, Alvin. (2009) "Williamson on Knowledge and Evidence." In *Williamson on Knowledge*, edited by Patrick Greenough, and Duncan Pritchard, Oxford: Oxford University Press, pp. 73-91.

Kelly, Thomas. (2008) "Evidence: Fundamental Concepts and the Phenomenal Conception of Evidence", *Philosophy Compass* 8, 5, pp. 933-955.

Littlejohn, Clayton. (2013) "No Evidence is False", *Acta Analytica* 28, 2, pp. 145-159.

Warfield, Ted A. (2005) "Knowledge from Falsehood", *Philosophical Perspectives* 19, 1, pp. 405-416.

Williamson, Timothy. (1997) "Knowledge as Evidence", *Mind* 106, 424, pp. 717-741.

__________________. (2000) *Knowledge and Its Limits*. Oxford University Press.

__________________. (2009) "Reply to Alvin Goldman", In *Williamson on Knowledge*, edited by Patrick Greenough, and Duncan Pritchard, Oxford: Oxford University Press, pp. 305-312.

University of Toronto
doug.roland.campbell@gmail.com

The Deontic Cycling Problem

William Simkulet

In his recent article “Deontic Cycling and the Structure of Commonsense Morality,” Tim Willenken argues that commonsense ethics allows for rational agents having both ranked reasons (A 〉 B, B 〉 C, and A 〉 C) and cyclical reasons (A 〈 B, B 〈 C, and A 〉 C). His goal is to show that not all plausible views are variations of consequentialism, as consequentialism requires ranked reasons. Here I argue apparent instances of deontic cycling in commonsense morality are the byproducts of an incomplete characterizations of the cases in question.

Keywords Deontic Cycling; Morality; Ethics; Trolley Case

The Deontic Cycling Problem

In "Deontic Cycling and the Structure of Commonsense Morality," Tim Willenken contends that "A range of extremely plausible moral principles turn out to generate "deontic cycling": sets of actions wherein I have stronger reason to do B than A, C than B, and A than C."[1] (545) He continues "... just about anything recognizable as commonsense morality generates deontic cycling." Rather than characterize apparent deontic cycling as a mistake in commonsense morality, Willenken contends that deontic cycling represents genuine insight into moral truth. For Willenken, deontic cycling is raised as a counterexample to consequentialist theories; but the scope of the objection is much further. The existence of genuine deontic cycles would constitute a counterexample to even James Rachel's modest metaethical view that the right thing to do in any given situation is the thing one has the best reasons for.[2]

A deontic cycle would be genuine if and only if it accurately reflects moral truth, or captures an actual law of ethics. Ethics is the branch of philosophy concerned with discovering what the right thing to do is in any given situation. The concept of deontic cycling poses a challenge to this enterprise, as the existence of genuine deontic cycling would mean that this enterprise fails. The problem of deontic cycling is that in a genuine case of deontic cycling, an action *x* may be both morally acceptable and unacceptable at the same time in the same way; but this

[1] See Tim Willenken, "Deontic Cycling and the Structure of Commonsense Morality", *Ethics* Vol. 122, No. 3, April 2012: 545-561.

[2] Although leading normative ethical theories may differ on what they believe constitutes a reason, most prominent theories, such as Kantianism and Utilitarianism, are consistent with Rachels' metaethics.

is analytically impossible. This paper is divided into two sections. In the first, I argue the notion of deontic cycling is incompatible with our commonsense moral beliefs. In the second, I show that Willenken's primary example of apparent deontic cycling fails to be a genuine case.

I.

The central goal of Willenken's paper is an attack on the view he (oddly) calls "compatibilism", the theory that consequentialism can be rendered consistent with commonsense morality; that commonsense moral beliefs are best understood in consequentialist terms. For example, a consequentialist might hold that the wrongness of lying is best understood in terms of undesirable consequences.[3] Consequentialist views require agents to rank possible actions by the strength of one's reasons to do them; Willenken argues that instances of deontic cycling show that moral reasons cannot be ranked.

Consequentialism holds that for any three morally inequivalent options A, B, and C, if B is morally preferable to A, and C is morally preferable to B, then C is morally preferable to A. (C > B > A) However, in an instance of genuine deontic cycling, B might be morally preferable to A, C morally preferable to B, and A morally preferable to C. (C > B, B >A, A > C) A genuine instance of deontic cycling would represent a counterexample to consequentialism. However, Willenken's target seems to be compatibilism, not consequentialism, and thus he isn't interested in whether commonsense morality generates as a whole genuine deontic cycles, rather if a deontic cycle is generated by one or more false, but

[3] For example, for a utilitarian, lying in a specific case would be wrong if it resulted in less overall utility, or happiness, than an alternative to lying.

commonsense moral beliefs, he believes it constitutes a counterexample to compatibilism.

Willenken's approach runs into two substantial problems. The first problem, I think, is a relatively minor problem - he has defined compatibilism too broadly. Compatibilists of the kind he discusses might explain deontic cycles in terms of situationally preferable consequences. For example, a compatibilist might contend that in an AB-situation (where one has to choose between A and B), B has relatively preferable consequences to A; in a BC-situation, C has relatively preferable consequences to B; and in an AC-situation, C has relatively preferable consequences to A. Although A, B, and C form a deontic cycle, in each situation the compatibilist claims to maximize the relevant consequences relative to the situation. Such compatibilists are even able to circumvent the real problem of deontic cycling, ABC-situations. An ABC-situation is a situation in which options A, B, and C are all simultaneously available. For Willenken, it seems as though each option - A, B, and C - would be both morally acceptable and morally unacceptable in the same way. However, this kind of compatibilist doesn't rank A, B, and C independent of a situation, as such in an ABC-situation whatever maximizes the preferable consequences turns on the situation itself, not any moral character of A, B, or C independently. Even if C is preferable to A in an AC situation, in an ABC situation A may be preferable to C.

The more substantial problem with Willenken's approach, however, is that although he argues that commonsense moral rules may generate deontic cycles, he doesn't argue that commonsense ethics abides deontic cycling. Our commonsense moral beliefs include beliefs about the nature of morality. For example, one commonsense moral belief is that there are no morally blind alleys; this is to say that one can never inadvertently put themselves in a position where they would be morally blameworthy for any choice they make.[4] A second, related, commonsense moral belief is the belief that ethics is complete; that there are moral rules governing any possible situation. A third commonsense moral belief is that although there may be both morally good and bad reasons for some actions, no action can be both morally acceptable and morally unacceptable at the same time in the same way.[5]

The existence of genuine, non-illusory deontic cycling would violate all three of these foundational commonsense moral beliefs. In a deontic cycle ABC-situation, every option is both morally acceptable and unacceptable, and although A, B, and C are morally inequivalent, it is a situation where ethics is not prescriptive. Willenken is silent on the issue of such foundational beliefs; but it seems as though he has three options - (i) reject that the foundational, metaethical beliefs described above are part of commonsense ethics, (ii) give up on the principle of non-contradiction, or (iii) embrace the view that some of our commonsense moral beliefs may be wrong. In cases of apparent contradiction between beliefs, one might engage in a reflective equilibrium to see which stay and which go. It strikes me that the foundational, metaethical nature of the beliefs described above give them the leg up over the commonsense moral rules he discusses.

[4] Michael Otsuka's *Principle of Avoidable Blame*, seems to capture this commonsense moral belief. See Michael Otsuka, 1998, "Incompatibilism and the Avoidability of Blame", *Ethics*, Vol. 108, No. 4: 685-701.

[5] Some actions, however, might be morally good and morally bad in different ways at the same time. For example, when your friend asks you whether you enjoyed your time together, you might lie to your friend to spare his feelings. This lie can be *prima facie* morally praiseworthy in that your intentions are to make your friend happy, yet it is also *prima facie* morally blameworthy in that your action is a violation of the trust between friends.

In an effort to demonstrate how deontic cycling is a result of our commonsense moral beliefs, Willenken constructs what he calls a “toy view” containing only two rules: (1) when faced with a choice between saving two boys, save the older boy, and (2) when faced with a choice between saving a boy and a girl, save the healthier of child. (549) When forced to choose between saving (u) saving a healthy young boy or (v) saving an unhealthy older boy, (1) requires him to save the older boy. When forced to choose between (v) saving a very unhealthy older boy, or (w) saving a moderately healthy girl, (2) forces him to save the girl. When forced to choose between (u) saving a very healthy younger boy, or (w) saving a moderately healthy girl, (2) requires he save the younger boy. Willenken asks us to choose between (u), (v), or (w). Here the toy view generates a deontic cycle, and thus there is no satisfactory answer. A deontic cycle is genuine only if the moral beliefs that generate it are true, but Willenken is under no illusions that the rules of the toy view are true.

Earlier I listed three foundational, metaethical commonsense moral beliefs; the second of which was that ethics must be complete. This is to say that a satisfactory normative ethical theory must tell you how to behave in any possible situation. The possibility of a genuine UVW-situation demonstrates that the toy view is incomplete - it fails to instruct the rules follower how to behave in that situation - and thus it is inconsistent with our commonsense moral beliefs. Of course we shouldn’t be at all surprised that the toy view is inconsistent with our foundational, commonsense moral beliefs - after all, it seems to violate a fourth foundational, commonsense moral belief - that ethics isn’t arbitrary. The two view arbitrarily identifies two values as morally relevant - age and health - but fails to offer an explanation as to why the values are valuable. As such, the toy view user lacks the tools they need to solve UVW-situations. Although Willenken contends that some of our (presumably non-arbitrary) commonsense moral beliefs generate deontic cycles, the toy view successfully illustrates that any genuine deontic cycle would demonstrate the incompleteness of ethics. After all, these commonsense moral beliefs would only tell us what things are valuable, but they fail to explain why they are valuable.

Independent of the incompleteness problem, Willenken's openness to the possibility of genuine deontic cycling meets with another problem - it requires an overly burdensome ontology. To paraphrase Occam's razor, when two theories offer the same explanatory value, the ontologically simpler of the two is to be preferred. Willenken's theory seems to be that at least some of our commonsense moral beliefs are independent, irreducible moral laws applicable to our lives that may lead to deontic cycles and the incompleteness of normative ethics. There are two issues with this view: First, if the view offered more explanatory value than the alternatives, it is so ontologically burdensome as to be unwieldy. We'd be embracing the existence of independent, ontologically distinct moral laws merely to explain apparent instances of deontic cycling.

Second, it's not at all clear that his theory does explain more than the average compatibilist theory. For many consequentialists and non-consequentialists, our commonsense normative ethical beliefs are *rules of thumb*, and not intended to be strictly applicable to each situation. Indeed, this seems to be how we actually intend many, if not most, commonsense moral rules to be used. Rather than our commonsense ethical beliefs being their own entities; consequentialists and non-consequentialists alike can hold that these rules of thumb are derived from a far smaller number of moral laws. Such a view better reflects how we use our commonsense moral laws, and has the virtue of not abandoning the completeness of ethics (and assuming the falsity of other foundational commonsense moral beliefs in the process).

Willenken seems to embrace the existence of deontic cycling solely because he believes no possible axiology will make consequentialism consistent with a view that includes genuine instances of deontic cycling, and because of this compatibilism is false. The price for this conclusion, though, appears to be a hobbling of ethics that flies in the face of commonsense moral beliefs more foundational than those he uses to support the existence of deontic cycling.

As Willenken has demonstrated with the toy-view, sets of beliefs can generate deontic cycling. The question that ethicists have to worry about is whether or not we have good reason to think that any sets of beliefs, commonsense or otherwise, that would generate deontic cycling actually capture true moral laws, rather than mere rules of thumb. However, even if one is committed to the position that the commonsense moral rules Willenken contends generate deontic cycles are genuine moral rules, rather than rules of thumb, these rules are still inconsistent with *prima facie* more foundational commonsense rules, such as the rule about the completeness of ethics. As such, either way one would be committed to the falsity of some of our commonsense moral beliefs. The question, then, is which set of rules - our foundational commonsense moral beliefs, or the subset of principles that generate deontic cycles - are we more committed to. The answer, I think, is clear.

II.

Willenken's primary example of deontic cycling comes from a series of trolley cases, where trolley cases are notorious for generating *prima* *facie* inconsistent sets of moral intuitions.[6] The apparently inconsistent moral intuitions generated by these cases, he contends, are actually the result of deontic cycling. Willenken generates his apparent deontic cycle with the following three cases:

Case 1:

There are two empty runaway trolleys, and you have the ability to stop one of these trolleys, but not the other. (Perhaps the switches you need to pull to stop both trains are too far apart to sprint to both in time.) The first trolley is barreling down a track that has five innocent people tied to it, the second is barreling down a track with two innocent people tied to it. You have two choices:

(x) Let five people die.

(y) Let two people die.

According to Willenken, commonsense morality dictates that (y) is preferable to (x), and that you ought to choose (y).

Case 2:

There is a single empty runaway trolley about to kill two people tied to a track. There is one way to stop the trolley before it kills both of these people: You can reposition one of these two people earlier on the track. If you do so, that person will die, but the other will live. You have

[6] Notable examples of trolley cases can be found in Philippa Foot, “The Problem of Abortion and the Doctrine of the Double Effect,” *Oxford Review*, Number 5, 1967: 5-15; Judith Jarvis Thomson, “Killing, Letting Die, and the Trolley Problem,” *The Monist*, 1976: 204-217; Judith Jarvis Thomson, “The Trolley Problem,” *Yale Law Journal*, 1985: 1395-1415; Peter Unger, *Living High and Letting Die*, Oxford: Oxford University Press, 1996; Francis Myrna Kamm, “Harming Some to Save Others,” *Philosophical Studies* 1989: 227-260; and William Simkulet, “Trolley Cases and Autonomy Violation,” Karios, Vol. 7, 35-48, 2013.

two choices:

(y) Let two people die.

(z) Kill one person.

Again, Willenken contends commonsense ethics contends (z) is preferable to (y), and that you ought to choose (z).

Case 3:

There is a single empty trolley traveling down a track with five people tied to it. The trolley is about to go under a bridge, then soon after it will hit the five people. There is only one way to stop the trolley before it kills all five people - you can push a relatively large person off of the bridge, and this person will fall to her death and derail the trolley. You have two choices:

(x) Let five people die.

(z) Kill one person.

Willenken contends that commonsense ethics dictates (x) is preferable to (z). He contends that commonsense ethics has generated a deontic cycle: (y) > (x), (z) > (y), and (x) > (z). For this to be a genuine deontic cycle, were there a case where you could choose from (x), (y), and (z), our commonsense moral beliefs would fail to be prescriptive. Consider the following case:

Case 4:

There is a single trolley barreling down tracks with five people tied to them. A mad villain, obsessed with proving the existence of deontic cycles, has tied you to a chair in the trolley control room. The villain explains that if you press button (x), all five people on the tracks will die, that if you press button (y), a switch will be turned, and the trolley will run over the first two people, but avoid the last three people. Finally, if you press button (z), a different switch will be turned, which will divert the trolley from its course - however in doing so, it will rip the first of the five people in half, killing her before the trolley has a chance to. You, thus, are confronted with the following choice:

(x) Let five people die.

(y) Let two people die.

(z) Kill one person.

It strikes me that in this situation our commonsense moral beliefs dictate that we ought to choose (z). But if there is a clear, commonsense moral choice in between (x), (y), and (z), then our commonsense moral beliefs are not generating a deontic cycle. How do we explain this?

Willenken's deontic cycle is created by equivocation between the various options listed in the cases, between (x^1) and (x^3); (y^1) and (y^2); and (z^2) and (z^3). The difference between these options is most apparent in the case of the last set. In case 2, (z^2) involves the killing of a person who would have died either way. James Rachels famously argues that, all else being equal, killing and letting die are morally equivalent, so in case 2 when you choose (z^2) you neither harm nor benefit that person in any substantial way (at worst, you shave a few moments off of his life to save another person).[7] In case 3, (z^3) involves the killing of an innocent person who would not have died unless you pushed him onto the tracks.[8]

[7] See James Rachels, "Active and Passive Euthanasia," *The New England Journal of Medicine*, Vol. 292, 1975: 78-80; "Killing and Starving to Death," *Philosophy*, Vol. 54, No. 208, 1979: 159-171.

[8] Oddly, case 3 in and of itself is a fairly effective argument against "compatibilism" and consequentialism, as it is an instance where in a choice between one life and five lives, our commonsense moral beliefs appear to show that one life is more valuable.

Even if you hold there is a morally relevant difference between killing and letting die, the killing of the person in case 2 is substantially morally different than the killing of the person in case 3, if for no other reason than the fact that you have no control over whether that person dies in case 2, but have total control over whether the person dies in case 3. Willenken's deontic cycle is an illusion by equivocation - he treats (z^2) and (z^3) as morally equivalent when they are not. Even if each of the other options were morally equivalent between cases, all he has shown is that (z^3) > (y) > (x) > (z^2). But (x^1) and (x^3) are different, where (x^1) is the option to save two people and let five other people die, while (x^3) is the option to let five people die rather than kill an innocent person. In any charitable reading (x^1) and (x^3), (y^1) and (y^2), and (z^2) and (z^3) are not morally equivalent to their counterpart, thus these three cases do not demonstrate even an apparent deontic cycle.[9]

Willenken recognizes that by distinguishing between the options in unlike cases, what he calls "fine-grained individuation", one can "can make deontic cycling disappear." (558) However, he says that this strategy is "is worryingly ad hoc, since our ordinary descriptions of actions... implicitly refer to the alternatives." This contention by Willenken is puzzling - descriptions are usually, by their nature, incomplete - they capture part of what is being described, but are not definite descriptors. Perhaps Willenken thinks that our ordinary action descriptions capture all of the morally relevant features of the action, but this would be absurd. For example, consider two actions of killing that are *prima facie* morally inequivalent:

Hostage Case 1:

[9] Stephen C. Making makes a similar argument in "Action Individuation and Deontic," *Ethics*, Vol. 123, No. 1, 129-136, 2012.

John, a police officer, is called to the scene where a violent escaped criminal is holding a hostage. John believes that the hostage's life is in danger, and that the criminal might escape to threaten other people, and that the best way to free the hostage is to shoot and kill the criminal. John shoots the criminal, intending to kill the criminal as a means to free the hostage. He succeeds, the criminal is killed by his shot, and the hostage is freed.

Hostage Case 2:

Joan, a police officer, is called to the scene where a violent escaped criminal is holding a hostage. Joan believes that the hostage's life is in danger, and that the criminal might escape to threaten other people, and that the best way to free the hostage is to shoot and kill the criminal. Joan also likes killing people. Joan shoots the criminal, taking this opportunity to satisfy her bloodlust in a way that will look like responsible police work. She succeeds, the criminal is killed by her shot, and the hostage is freed.

In these cases John and Joan both act to kill the criminal, but John is morally praiseworthy for his action, while Joan is not. Of course we are not often privy to the inherently private mental states of others, so when we witness a police officer shooting a hostage in a case like this, we can only judge them with incomplete information. The difference between John and Joan is the moral intention with which they act. However, both actions can be described as "killings," as such our ordinary action descriptions fail to capture all of the relevant moral features of their actions.

As such it strikes me that a proper analysis of Willenken's cases requires a full account of the intentions with which the agents act. For example, in case (1), the choice isn't between (x^1) - let five people die

- and (y^1) - let two people die -, it's a choice between (w^1) let all seven people die, (x^{1a}) act to save the two people first, then try to save the five people, (y^{1a}) act to save the five people first, then try to save the two people, (x^{1b}) act to save the two people first so as to appear virtuous, and pretend to try to reach the last five, but purposely fail so you get to enjoy watching five people die, so forth and so on. Willenken treats the options in case 1 as if the intentions of the agent in question are morally irrelevant, and the outcome is certain; but neither is the case.

Were one faced with the decision in case 1, it strikes me that the right choice is (y^{1a}) - you try your best to save both sets of people, starting with the larger set. It may be impossible to save both sets, but to not try to save both sets is, I think, uncontroversially morally abhorrent. Suppose that you were to watch someone race towards the first track, and throw the level as hard as they could so as to save the five people imperiled by the first trolley, then sit back leisurely as the second trolley runs over two people. I imagine we'd judge such a person morally despicable - if there is even the slightest chance you could save the second set of people, commonsense morality dictates that you try.

Conclusion:

The apparent instance of deontic cycling between cases 1-3 is generated by equivocation between unlike expected outcomes. Case 3 represents a genuine moral dilemma - we are committed to the proposition that killing and letting die, all else being equal, are morally equivalent, but that when forced with the choice between killing an innocent man and letting five innocent people die, our commonsense moral intuition seems to commit us to choosing the latter.

This case draws our attention to a genuine inconsistency in our commonsense moral beliefs, but Willenken denies this, instead contending that our commonsense moral beliefs are consistent, but incomplete - cobbled together from disparate irreducible moral principle that each capture a different moral truth. This move renders ethics incomplete and bloats our ontological commitments with no discernible benefit. This is not a move worth making.

References

Foot, Philippa. (1967) “The Problem of Abortion and the Doctrine of Double Effect”, *Oxford Review* 5, pp. 5-15.

Kamm, Francis Myma. (1989) “Hamming Some to Save Others”, *Philosophical Studies* 57, pp. 227-260.

Makin, Sthephen C. (2012) “Action Individuation and Deontic Cycling”, *Ethics* 123(1), pp.129-136.

Otsuka, Michael. (1998) “Incompatibilism and the Avoidability of Blame”, *Ethics* 108(4), pp. 685-701.

Rachels, James. (1975) “Active and Passive Euthanasia”, *The New England Journal of Medicine* 292, pp. 78-80.

____________. (1979) “Killing and Starving to Death”, *Philosophy* 54(208), pp. 159-171.

Simkulet, William, (2013) “Trolley Cases and Autonomy Violation”, *Kairos* 7, pp.35-48.

Thomson, Judith Jarvis. (1976) “Killing, Letting Die, and the Trolley Problem”, *The Monist* 59(2), pp. 204-217.

___________________. (1985) “The Trolley Problem”, *Yale Law Journal* 94, pp. 1395-1415.

Unger, Peter. (1996) *Living High and Letting Die*, Oxford: Oxford University Press.

Willenken, Tim. (2012) “Deontic Cycling and the Structure of Commonsense Morality”, *Ethics* 122(3), pp. 545-561.

University of Wisconsin, Marshfield/Wood County
simkuletwm@yahoo.com

Habermas and Mouffe on the Question of Rationalization

Aidan M. Sprague-Rice

Chantalle Mouffe has made two interesting arguments against the desirability of social rationalization. The first is that the political project of supporting the development of a rationalized society produces a motivation challenge that renders the project difficult to accomplish. Social movements concerned with bringing about a more just world, then, ought not to understand their project as primarily directed toward making the world more rational. The second is that, if a fully rationalized society ever were achieved, it would be prone to violence. In this paper I use the theory of communicative action, developed by Jürgen Habermas, to rebut Mouffe's claims. When we theorize rationality as communicative action instead of instrumental rationality, we come to see that neither the motivation challenge nor the argument from violence can stand as compelling arguments against the desirability of social rationalization.

Keywords Jürgen Habermas, Chantalle Mouffe, Rationalization, Communicative Action, Instrumental Rationality, Violence

In his central theoretical work, the *Theory of Communicative Action* (hereafter *TCA*), Habermas is concerned with defending the desirability of social rationalization against early skeptics of the Enlightenment political project (most notably Adorno and Horkheimer). Since the publication of *TCA*, new critics of rationalization have emerged. One of the most tenacious and best received of these has been Chantalle Mouffe. In her recent book, *The Democratic Paradox*, Mouffe makes two anti-rationalization claims: (1) for practical reasons, rationalization is not an ideal that political movements concerned with producing a better world should take up, and (2) the achievement of a more rational society would have undesirable, violent, consequences.

In support of her first claim, Mouffe argues that political movements which take up rationalization as a guiding ideal will produce ambivalent attitudes about politics among their members. Taking rationalization as a primary goal will thus be self-defeating for political groups because doing so will reduce the willingness of their members to undertake the political action that would be necessary to make their societies more rational.

In support of her second claim, Mouffe argues that the achievement of a rational society would involve bringing about a world in which political groups with views distinct from those of the powerful would be vulnerable to oppression. This is because the powerful groups in a rationalized society could claim a transcendental justification for their political views such that alternatives could easily and effectively be framed as incorrect and socially de-stabilizing - hence things to be resisted, even at the cost of violence.

In what follows I will provide a reconstruction of Mouffe’s two arguments. I will then show that they presuppose a traditional, instrumentalist, account of rationality. Once this is established, I will show that the pro-rationalization political project appears in a much different light if we adopt an alternative account of rationality, for instance the communicative account developed by Habermas in *TCA*. I will show that, if we take Habermas to be right about rationality, Mouffe's two concerns about rationalization will be unfounded. This suggests that the prospect of an increasingly rationalized society may indeed be a worthy goal for political groups. Whether or not this is the case will ultimately depend upon which theory of rationality – the instrumentalist account favored by Mouffe or the communicative account favored by Habermas – is correct (assuming that there is not some third, better, account available).

Before proceeding to Mouffe's arguments, it is prudent to define some key terms. First of all, I distinguish between pro- and anti- rationalization theorists. Pro-rationalization theorists think that there is something good about the prospect of ordering political life in a rational way. They thus desire that social life should come to be characterized by procedures for formulating and resolving political questions rationally. They support the creation of institutions that would make this possible and the adoption of rational solutions to particular political problems.

Anti-rationalization theorists, on the other hand, think that there is something problematic about the prospect of a world in which political questions are formulated and resolved rationally. They thus oppose (at least to some extent) the creation of institutions which would make this possible and resist (at least to some extent) the adoption of rational solutions to particular political problems.

Finally, I take political questions to be questions about how members of a community ought to identify and resolve disputes that arise within their community.

1. Mouffe

Now that I have specified the meaning of these basic terms, let us consider Mouffe as a representative for contemporary anti-rationalization theorists. I can discern two interesting anti-rationalization arguments in Mouffe's recent work, *The Democratic Paradox*. I will call the first The Ambivalence Argument, and the second The Violence Argument.

1.1 The Ambivalence Argument

The Ambivalence Argument is meant to show that accepting the ideal of rationalization, and thus advocating for rational solutions to political questions, leads to the development of a widespread malaise about political action which inhibits the realization of a rational society. If Mouffe can support this conclusion, then she will have shown that taking rationalization as a guiding political ideal is self-defeating (since accepting the ideal inhibits making progress towards its realization).

To support her claim, Mouffe must provide us with (1) an account of why people are motivated to participate in political activity and (2) an explanation of why adopting the project of creating a rational world would interfere with this motivation.

For Mouffe, modernity is characterized by dual political commitments that individual people, moderns, take on through socialization. The first is a commitment to democracy. The second is a commitment to liberalism (TDP, 2-3).

Mouffe describes the commitment to democracy as (among other things) a commitment to the rule of the people over themselves ("identity between governing and governed") and to the establishment of equal standing for all members of the people (TDP, 2-3). She describes liberalism, on the other hand, as (among other things) a commitment to the *universal* enjoyment of basic human rights and personal liberty (TDP, 2-3).

Now, on Mouffe's account, the commitment to democracy is so fundamental for moderns that it is (perhaps uniquely) capable of serving as the basis for prolonged and transformative political action (TDP, 4). That is to say, we moderns have so internalized a commitment to democracy that, when we feel that our society is not democratic, we will work together, sometimes under the threat of serious danger, to make it more democratic.

Mouffe also holds that pro-democratic political action depends upon the ability of a group of individuals to see themselves as members of a specific people (over against other peoples). Only then can they think of themselves as collectively being governed by some other person or group, hence as living in a world that is not yet democratic, such that political action to make the world more democratic could be called for.

Now it is just this requirement for pro-democratic political action, that individuals be able to see themselves as members of a specific people, which Mouffe thinks that people who understand themselves as committed to the development of a rational society cannot fulfill. Mouffe's argument in favor of this claim basically reduces to two premises: (1) in order to understand oneself as a member of a people, one must understand some other person or group to **not** be a member of that people, and (2) once we have accepted rationality as our guiding political ideal, we will not be able to consistently formulate a distinction between ourselves and other people such that we could understand ourselves as members of a specific people.

Mouffe uses Carl Scmitt's analysis of liberal universalism to establish her first claim (although she disavows his ultimate political conclusions) [TDP, 8]. Schmitt famously argued that the self-conscious constitution of a people qua people could not be accomplished absent the assignation of other-ness to some person or group. Mouffe takes him to have a hold of a deep realization here with major political implications. Perhaps this is true, although I am not sure that it is. However, despite my skepticism of the implications for political theory, I will accept Mouffe's first premise because I think that, even if it is true, it is not a major problem for pro-rationalization political movements.

Mouffe's second claim is the more important one for my purposes in this paper. She claims to be able to show that, once we accept rationality as our political goal and our standard for adjudicating political disputes, we will no longer be able to make discriminations between people such that a pro-democratic political movement could develop. Why does she believe this to be true?

Mouffe does not provide a formal statement of her account of rationality in *TDP*, so we cannot be completely sure about which features of rationality she thinks prevent us from generating a robust notion of the people to which we belong. Nevertheless, if we take Mouffe's analysis to be operating with the instrumentalist theory of rationality that dominates much contemporary analytic political philosophy (one of Mouffe's primary targets in *TDP*) in the background, we can understand why she thinks as she does.

When we accept the instrumentalist account, we understand rationality to be a procedure that is *exclusively* suited to help us come to true beliefs about the (objective) world and to learn how to successfully intervene in it. As Hume formulated it in *An Enquiry Concerning Human*

Understanding, thinking rationally can give us insight only into "relations of ideas" and "matters of fact" (*Enquiry*, 108). So, if the instrumentalist account is true, rationality *cannot* ground value claims. This is the realization behind Hume's is-ought distinction.

One implication of the truth of the instrumentalist account of rationality is thus that thinking in an exclusively rational way prevents us from identifying any trait or value that some person might possess as the proper criterion for determining who ought to be understood as equal to us (members of our group or of our people) and who ought to be denied this status. About moral questions like these, those who are committed to generating their beliefs through instrumentally rational procedures must remain silent, because they simply have no way available to them to ground the claim that one set of characteristics is the one that we *should* adopt to demarcate different peoples. But if rationality is of no help to us in deciding these questions, then it seems that Mouffe is right to hold that pro-rationalization theorists, since they are committed to answering political questions rationally, cannot ground the distinction between peoples that is necessary to mobilize the distinctively modern commitment to democracy as motivation to participate in pro-rationalization political action. In this case pro-rationalization theorists would indeed damage the prospects of their project by formulating and accepting it as an ideal.

1.2 The Violence Argument

Alongside The Ambivalence Argument, Mouffe offers The Violence Argument against pro-rationalization theorists. The Violence Argument is supposed to show that, if a rational society ever were achieved, it would be characterized by violent oppression of political dissidents. Since this is would be an obviously horrible outcome, Mouffe is entitled to reject rationalization as a desirable political ideal.

Unlike The Ambivalence Argument, The Violence Argument is somewhat easy to unpack. This is partially due to the fact that it is a familiar argument of anti-rationalization theorists (see, for instance, Berlin, 179-180) that Mouffe has merely reformulated in *TDP*. My summary of it will therefore be somewhat brief.

The first claim in the argument is that, in a rationalized society, political institutions and laws would be understood, by those who accept rationality as the proper way to order political affairs, to be correct (because they are rationally justifiable). The second claim is that, if a state of affairs is rationally justifiable, then it cannot lose its legitimacy in the eyes of those who accept rationality as the proper method for ordering political life. This is because rationality is understood, by those who accept it as the proper method to be used in ordering political affairs, to be an ahistorical method, hence one which will always endorse the same conclusions (regardless of the historical moment in which it is deployed). The third major claim is that, when people confront the possibility of an apparently incorrect mode of organizing political life becoming dominant in their society, they will strongly resist it, resorting to violence if they have to, because people take the correct organization of political life to be a matter of first importance. Once all this is established, we can see why Mouffe thinks that political dissent would be subject to violence in a rationalized society. Because dissent from the established institutions would necessarily appear to be irrational, hence incorrect, we can expect that it would provoke violent reactions because it would appear as a serious threat to the health and stability of society.

In her attack on Rawls' pro-rationalization arguments, Mouffe makes it clear that she has something very much like the above reconstruction in mind.

"Rawls seems to believe that whereas rational agreement among comprehensive moral, religious, and philosophical doctrines is impossible, in the political domain such an agreement can be reached. Once the controversial doctrines have been relegated to the sphere of the private, it is possible, in his view, to establish in the public sphere a type of consensus grounded on Reason (with its two sides: the rational and the reasonable). *This is a consensus that it would be illegitimate to put into question once it has been reached, and the only possibility of destabilization would be an attack from the outside by 'unreasonable' forces. This implies that when a well-ordered society has been achieved, those who take part in the overlapping consensus should have no right to question the existing arrangements, since they embody the principle of justice. If somebody does not comply, it must be due to 'irrationality' or 'unreasonableness'.*" (TDP, 28-29, emphasis mine).

2. Habermas's Alternative to Mouffe

Now that we have summarized Mouffe's two objections to rationalization, we are in a position to comment on the theory of rationality that they presuppose. What has emerged is that Mouffe's account of rationality is a traditional one in two respects. First, Mouffe is operating with something like an instrumentalist account of rationality (i.e. she is following Hume insofar as she believes that rationality is of no use to us when our goal is to provide a strong grounding for ought-claims). Second, she retains an a-historical concept of rationality. That is, she does not accept that the set of claims which can be rationally justified vary at different historical moments. We will see now, turning to Habermas, that it may well be that neither of these characteristics are obviously part of a strong theory of rationality.

In the *Theory of Communicative Action*, Habermas develops an anti-foundationalist, yet (weakly) universalist, account of rationality – the communicative account of rationality. Here I will unpack the communicative account to show that, if it is correct, Mouffe's political anti-rationalization arguments must fail. I will argue that, if Habermas's account is correct, the Violence Argument will fail because anti-foundationalist accounts of rationality prohibit the traditional imputation of transcendental justification to rationally-justified claims such that Mouffe's concern about the ease that the powerful would have in branding challenges to their positions as dangerous (because irrational) will be unfounded. It will also fail because, from the perspective of an individual socialized into a communicatively rational society, those who non-violently resist dominant political perspectives must be respected insofar as their resistance amounts to an (at least potential) increase in the critical power of a community concerned with discovering the truth about some question. I will also show that the Ambivalence Argument will fail because communicative rationality, insofar as it involves truth-directedness, involves those participating in it in moral commitments that are robust enough to identify (always provisionally) others against whom pro-democratic energies could be marshalled (always non-violently) as part of the pro-rationalization project. I will defend these claims in turn.

2.1 Habermas's Anti-Foundationalism and the Violence Argument

I will begin with a discussion of Habermas's anti-foundationalism. This requires that I say a few words about foundationalism. I understand ‘foundationalism’ to refer to the claim that thinking rationally allows us to come to an understanding of being itself (i.e. absolute reality). Foundationalists believe that our beliefs can be known to be true reflections of absolute reality when they follow from a specific method of belief-generation, rationality. Rational justification ensures that our beliefs are true reflections of absolute reality because it is constrained by (or attuned to) the features of absolute reality itself. By being able to gain rational justification, then, our beliefs prove that they are responsive to, in some way reflections of, being-itself. Insofar as we are committed to the possession of true beliefs (and we might be so committed for a variety of reasons, for instance because we want to effectively intervene in the world), then, claims that are truly rationally founded compel belief for foundationalists.

Habermas announces his anti-foundationalism in the first chapter of TCA. There he describes the foundationalist project as having broken down (TCA, 2-3). Behind his rejection of foundationalism is a basic acceptance of the pragmatic tradition’s historicization of rationality. With the pragmatists, Habermas views rationality as a belief-generating and action-coordinating social mechanism which has evolved in response to challenges faced by human communities over time. The act of thinking rationally is thus not to be thought of as putting us in a position to produce transcendentally justified conclusions about the nature of being itself. We are not justified in believing rationality to be capable of producing such beliefs because we understand it to be itself acceptable only so far as it continues to enable us to meet the challenges we face from day to day better than available alternatives.

Now it is just this realization – that rationality is a historically contingent phenomenon – which Mouffe thinks will guard against the development of social oppression. She claims that, once we recognize that rationality is not suited to the task of generating transcendental knowledge about being itself, we will stop thinking that the discovery of the other's non-rationality is a good reason to react violently against the other. She refers to this realization as an acknowledgment of the limits of reason, one which it is crucial that we come to terms with because it will produce in us a sort of irony about rationality that will counteract the temptation to brand non-rational people as dangerously incorrect and hence rightly subject to paternalistic (perhaps even violent) intervention by rational people (TDP, 133).

It would seem to be the case, then, that Mouffe ought to exempt pro-rationalization theorists who are working with a theory of rationality like Habermas's from the Violence Argument. That is, she ought to be clear that the Violence Argument applies only to pro-rationalization theorists who are also foundationalists. However, she fails to do this in TDP. There she includes Habermas among the theorists to which the argument is supposed to apply.

Mouffe's failure to appreciate Habermas's anti-foundationalism notwithstanding, her attempt to ground protection of non-rational others from oppression by appeal to the significance of the discovery of rationality's contingent historical nature ultimately fails because it is coupled with a rejection of rationalization and thus a rejection of the very inferential principles which must be in operation if rationality's historical contingency is to be consistently thought to imply the rightness of a prohibition on violent oppression of others because of their non-rationality. We can appreciate this point if we think about what a non-rationalized political sphere would be like. There something other than the rules of inference (e.g. the principle of non-contradiction) which Mouffe thinks are definitive of rationality would be in operation. But, in that case, we could not predict what an appreciation of rationality's historical pedigree would lead us to do or believe. Indeed, the moderating impact on our actions brought on by the discovery of rationality's contingency depends upon our combination of the discovery of its contingency with more general moral or political principles in the form of an argument which produces the conclusion that we must not take the commitments which result from rational reflection to be good reasons to act violently. And the formation of such an argument will only occur if we have already accepted basic logical principles, like the principle of non-contradiction, which we would not be compelled to accept in Mouffe's non-rationalized society.

Consider the following unpacking of the discovery of rationality's contingency as an example to help make this clear. We might utilize a principle of caution about undertaking actions from the perspective of a belief-generating procedure which is subject to change in unpacking the importance of rationality's contingency. In this case we might hold:

> 1.) If we might replace our belief-generating procedures over time, then we should be wary of acting on the beliefs which they produce when such action would have irrevocable consequences (because we might later, after the procedure has changed, wish that we had not acted in this way).
> 2.) Rationality is a belief-generating procedure which we might eventually replace.
> 3.) So we should be wary of acting on rationally justified beliefs when such action would have irrevocable consequences (e.g. when such action would be seriously violent).

We only feel compelled to accept the conclusion of this argument if we begin from the belief that contradictions cannot be true. This tells us that the negation of (3) cannot be true if (1) and (2) are true. But Mouffe cannot depend on this basic logical commitment to the impossibility of true contradictions, because a widespread commitment to it in the realm of political discourse is something that she would eliminate by rejecting social rationalization.

Now, if we were to generate political commitments using something other than the familiar principles of logic we would need to specify which principles were going to guide our interpretation of the implications of our premises (beliefs) before making a guess about the implications of the realization of rationality's contingency. Mouffe, however, does not specify what would replace rationality for us in her non-rationalized political sphere. She merely claims that recognition of rationality's contingency triggers a respect for otherness. We can, however, easily imagine alternatives to this happy outcome. The contingency of belief-formation procedures could be taken, for instance, to be a good reason to reject all such procedures and to merely act in accordance with one's own immediate desires (because there is no standard according to which beliefs ought to be set and hence no reason to reflect or generate beliefs at all). In the absence of a specification of the inferential principles which Mouffe thinks will characterize a non-rationalized society, then, she has not made good on the claim that the recognition of rationality's contingency will lead to an irony about rationality which will protect difference. Only a pro-rationalization theorist, like Habermas, who is in a position to rationally unpack rationality's own limitedness, can be expected to consistently draw from the realization of rationality's contingency the conclusion that we ought to be non-violent in our interactions with non-rational people. Thus Habermas's communicative account of rationality can meet the Violence Objection, but it is not clear that Mouffe's political anti-rationalization proposal can be known to be non-violent.

2.2 Rationality and Moral Content in Habermas

Putting aside the issue of the implications of his anti-foundationalism, there are other good reasons to hold that accepting Habermas's communicative theory of rationality could put pro-rationalization theorists in a position to meet Mouffe's objections. To see how, we must reflect on the way that individuals socialized into communicative rationality would use language to organize their relationships with one another.

In TCA, Habermas forwards the (empirical) hypothesis that there are general features which characterize all rational speech-acts (i.e. all attempts to use language rationally). Each speech-act contains (1) a claim to describe the world as it actually is (a truth claim), (2) a claim to be an appropriate utterance relative to other people (an intersubjective, normative, claim), and (3) a claim to be an authentic expression (a subjective claim). To make this concrete, take the speech-act: "You should lighten up and relax." On Habermas's account, this speech-act contains at least one truth claim, one normative claim, and one subjective claim. For instance, we might interpret the speaker as claiming (1) that the subject to whom the speech-act is directed is behaving anxiously (a truth claim), (2) that the speech-act itself is an appropriate attempt to change the behavior of the subject to whom it is directed (a normative claim), and (3) that the speaker who performs the initial speech act is behaving earnestly, that he is not, for instance, jesting with the person to whom the speech-act is directed (a subjective claim).

Now, the fact that we use language to relate to one another in this way indicates, according to Habermas, that we are operating with basic world-concepts – general beliefs about the structure of the worlds we inhabit – that make sense of our utterances. In the attempt to redeem the validity claim to truth, for instance, what we are doing is attempting to show that our claims about the world match up with the way that the world actually is. In doing so we presuppose that we confront a world of things which really are just one way at any given time such that our beliefs could in fact match up with them (because there is something, the way that the world actually is at this moment, for our beliefs to match up with). The background world-concept (of a world of things that are just one way) operating here is called, by Habermas, the objective world-concept. Something similar is required in order for the attempt to redeem the normative validity claim to make sense. In this case we must presuppose that we exist in relation to other subjects who are like us at least insofar as we and they are both capable of generating and following rules of interaction. Absent this possibility, it doesn't make sense to talk about whether or not an interaction is appropriate because a rule-governed interaction itself is impossible. Habermas calls this the intersubjective world-concept. And something similar is happening, too, when we make the claim that an utterance is sincere. Here we presuppose that it could be insincere, hence that we contain an internal world where authentic beliefs could be formulated by us and then concealed. Habermas refers to this as the subjective world-concept.

Now, crucially, from the perspective of individuals socialized into a rational way of life, the relevant world-concepts do more than make sense of the validity-claims that we make when we perform speech acts. They also, in the final analysis, structure our attempts to redeem our validity-claims to other people. To see how, consider the example of a dispute about whether or not something is the case that arises between two people who are ready to settle it rationally (i.e. who are willing to use reasons, as opposed to force, to generate a common belief). To determine who is correct in this case, the different speakers will attempt to trace implication relationships between (1) their proposed beliefs and (2) shared beliefs about the way that the world is such that denying their proposed beliefs would involve a contradiction. If a relationship of implication really does hold between some proposed belief and beliefs accepted by all the other parties to the dispute, then the proposed belief must be accepted because all the parties to the dispute accept the objective world-concept, that the world is just one way at any given time, from which the basic principles of logic (e.g. the principle of non-contradiction and the Law of Excluded Middle) can be read off. The objective world-concept, then, functions to underwrite rules for determining when an argument is compelling and when not. It thus enables speakers who disagree to come to agreement without resorting to violence, by using reasons instead.

But Habermas's world-concepts do more than ground a commitment to the basic principles of logic. They also commit us to developing tolerance and respect for difference in our daily interactions with other people and in our social institutions. To see why, consider the form of reflection we would have to adopt if we were truly interested in settling disputed validity claims, like the claim to truth, with reasons. In this case it would not be enough to merely accept the basic logical principles which help us to discriminate between correct and incorrect inferences. Because we know that we are epistemically limited creatures (e.g. capable of drawing bad inferences and thinking that they are good, capable of deluding ourselves about what we know) we would also need to be committed to engaging in critical dialogue with all willing conversation partners directed at discovering how things really are in the world. Only in this case would we have applied the full possible critical scrutiny to our own thinking such that the mistakes we might be making could be recognized. And we must be committed to doing this because we believe that the world is some way (just one way), such that we could be mistaken about it.

In other words, then, the fact that rationality involves forming beliefs which we hope will actually mirror the features that the world has, coupled with the fact that we know that we are capable of making mistakes and not knowing that we are doing so, means that we must be willing to engage with all other people (who are willing to engage with us) about the matter that we are interested in when our goal is to discover the truth about something. These others can contribute to our truth-directed pursuit their own critical competence, which diminishes the chance that we will ultimately believe something to be true for bad reasons (i.e. the chance that we will make a mistake in formulating our beliefs). But this means, further, that being committed to rationality involves being committed to a specific set of moral principles. To be rational is to be committed to treating others with respect (such that they actually can participate in a critical, truth-directed, dialogue) and providing them with (at least) the basic material resources necessary to use their critical thinking skills in a truth-directed discourse. And this means, further, that being rational requires that we be committed to developing political institutions and arrangements which enable people to become full participants in critical-rational dialogue. This means developing a society in which dissenting views are allowed to develop and are treated with respect. Contra Hume and Mouffe, then, Habermas shows that rationality actually does have moral content insofar as thinking rationally involves making a commitment to an ethic of respect and care for the other.

3. Habermas's Theory of Rationality as an Answer to Mouffe

Now, if Habermas's theory (as I have reconstructed it here) is basically right-headed, there are important implications to the ability to connect up rationality with an ethic of respect and care. First, we can see that this means that, counter to Mouffe's worry that a rationalized world would be one in which political opposition groups would be oppressed; different perspectives would be welcome and protected. This is because they would be seen as part of an ongoing conversation about the appropriate way to govern our lives in common. The Violence Argument thus fails (as a good objection to a pro-rationalization theorist like Habermas) for a second time.

To repeat, this is because being committed to rationality already means being committed to a moral perspective from which institutions or practices which fail to provide individuals with basic resources or respect cannot be seen as justifiable. Rationality thus provides us with a strong moral criterion. This is something that Mouffe does not notice in her development of the Violence Argument.

Second, we can see that Mouffe's concern that pro-rationalization political movements will not be able to ground a strong self-identity such that they can marshal the modern commitment to democracy to generate motivation for political action also fails. The reason is that the prohibition on violence that we accept as part of being rational can be used as a measuring stick to determine who is part of the pro-rationalization movement and who is not. Those who resort to using violence to settle political disputes (and there have always been such groups) will mark themselves as others against which pro-rationalization theorists can define themselves, thus overcoming Mouffe’s worry about the impossibility of grounding a shared identity for members of the pro-rationalization political project. The Ambivalence Argument, then, also fails when applied to Habermas’s account.

Whether or not Habermas’s account of rationality is ultimately tenable is a question which is too involved to approach in this paper. What I have shown here, instead, is that if it is correct then the project of rationalization is not doomed to failure for the reasons which Mouffe suggests.

References

Berlin, Isaiah. (2002) *Liberty: Incorporating Four Essays on Liberty*, Oxford: Oxford UP,. Print.

Habermas, Jurgen. (1984) *The Theory of Communicative Action, Volume 1, Reason and the Rationalization of Society*. Boston: Beacon.

Hume, David. (1999) *An Enquiry Concerning Human Understanding*. in Tom L. Beauchamp. (ed.), Oxford: Oxford UP.

Mouffe, Chantal. (2009) *The Democratic Paradox*. London: Verso,.

Department of Philosophy, Michigan State University
spragu55@msu.edu

철학적 분석 31 (2014) pp.87-121

Against Inferential Reliabilism: Making Origins Matter More

Peter J. Graham

Reliability theories of epistemic justification face three main objections: the generality problem, the demon-world (or brain-in-a-vat) counterexample, and the clairvoyant-powers counterexample. In *Perception and Basic Beliefs*(Oxford 2009), Jack Lyons defends reliabilism at length against the clairvoyant powers case. He argues that the problem arises due to a laxity about the category of basic beliefs, and the difference between inferential and non-inferential justification. Lyons argues reliabilists must pay more attention to architecture. I argue this isn't necessarily so. What really matters for understanding and solving the case involves paying closer attention to the origins of our belief forming capacities, both inferential and non-inferential. Reliabilists should make origins matter more.

Keywords Epistemic justification, evidentialism, reliabilism, inferential reliabilism, clairvoyance, Jack Lyons, hopeful monsters

John wakes up, opens his eyes, and looks out the window. Lo and behold, another sunny day. His beliefs (it's morning, it's sunny outside, it's sunny outside again, it's going to a pleasant afternoon) are surely all "justified."

Being justified in general means *being in the right*. This involves meeting some standard or norm for correctness. *Being justified* in epistemology means being in the right vis-à-vis the goal of believing truth and avoiding error. A justified belief then meets a standard or norm understood in terms of promoting truth and avoiding error.

Many traditional epistemologists connect justified belief to the individual's ability to justify her belief. This view has fallen on hard times for it overly narrows the scope of justified beliefs. Small children and many non-human animals have justified beliefs. But they lack the capacity to critically reason in support of their beliefs. John's beliefs may be justified even if he's only four years old.

There's no real doubt *whether* John's beliefs are justified. But there's a real philosophical issue accounting for *why* they are justified. Are they justified because they are based on *conscious* sensory perceptions, *conscious* episodes of propositions seeming to be true, and *conscious* episodes of one set of beliefs *consciously seeming* to support another?

"Experientialists" (evidentialists, mentalists, dogmatists, phenomenal conservatives) say yes, indeed, that's why they are justified. Our conscious, sensory perceptions, among other conscious states and events, explain why garden-variety beliefs based on perception, stored in memory, and extended through reasoning are justified. There is something about conscious, sensory perceptions and other conscious states and events that explain why they justify. Basing beliefs on "experiential evidence" is the standard by which beliefs are justified. Experientialists believe in the epistemological power of consciousness.

"Reliabilists" say no. The matter of fact (unconditional) reliability of perception — the fact that perception produces mostly true beliefs in our ordinary circumstances — explains why perceptual beliefs are (unconditionally) justified. The matter of fact (conditional) reliability of memory — the fact that memory reliably preserves beliefs previously formed in some other way — explains why beliefs stored in memory are (conditionally) justified. The matter of fact (conditional) reliability of reasoning — the fact that reasoning reliably transitions from true premises to true conclusions — explains why reasoned beliefs are (conditionally) justified. Getting things reliably right is the standard by which beliefs are justified.

Experientialists agree that perception, memory, and reasoning are reliable. They also typically agree that reliability matters to *knowledge*. But they reject the idea that the reliability of these processes explains why their outputs — perceptual beliefs, memory beliefs, and inferential beliefs — are *justified*. Experientialists see justification as a good "internal" fit between our beliefs and our conscious seemings or experiences. Reliabilists, on the other hand, see justification as a good "external" fit between our belief-forming processes and the states of the world our beliefs represent.[1]

Enter Jack Lyons. If anyone is a dyed-in-the-wool reliabilist, Lyons is. And if anyone can't stand experientialism, Lyons can't. In *Perception and Basic Beliefs: Zombies, Modules, and the Problem of the External World*(Oxford, 2009), Lyons takes on two main tasks: wage war against experientialism and develop and defend reliabilism.

[1] I discuss these contrasts in more detail in my 2011a and forthcoming-a.

Let us assume, for the sake of argument, that Lyons is right about experientialism. Let's also assume, for the sake of argument, that reliabilism is the only plausible alternative. Does it follow that our work is done? Hardly. For Lyons and I agree (along with many others) that reliabilism faces a number of problems to overcome. Even if experientialism is dead in the water, reliabilists still have work to do.

Lyons takes up the clairvoyant-powers problem in his book and the new evil-demon problem in a follow-up essay (Lyons 2012). In this paper I'll critically engage, from a reliabilist perspective, Lyons treatment of the clairvoyant-powers case, putting experientialism entirely to one side.

Lyons argues for two theses. First, that the clairvoyant powers case arises due to an unfortunate "laxity" among reliabilists about "inferential justification." Second, that in order to solve the problem, besides paying more attention to the difference between inferential and non-inferential justification, we need to add an "etiological" or right "origins" condition on justified belief.

I shall argue for two counterpart theses. First, the problematic clairvoyance cases really arise because they've got the wrong kind of origin, not because reliabilists have been lax about the inferential vs. non-inferential distinction. Second, because the problems arise due to wrong origins, Lyons needs to say considerably more about why the origins he cites really matter. "Laxity" about "inferential justification" is a red herring; what really matters is unacceptable laxity about origins, and Lyons's own discussion of origins, unfortunately, is unacceptably lax. Reliabilism needs a good account of why origins matter, and Lyons fails to provide one.

I. The Trouble with Simple Reliabilism

In his ‘Response to Critics’ Lyons indulges in some useful autobiography:

> This project started off in my mind as a way of solving a problem for reliabilist theories of justification, namely, their unacceptably lax treatment o f…inferential justification. Clairvoyance cases are just the tip of an iceberg: it seems undeniable to me that some beliefs require argument, that they require inferential, or doxastic support, if they are to be justified. “Simple reliabilism” holds that reliability is sufficient for *prima facie* justification, thus, in essence, denying that any belief requires inferential support. But take any hard-won item of science or philosophy: the belief that reliabilism is true, that bats are more closely related to primates than to rodents, that the moon is 2178 miles in diameter, and so on. There are many more: my belief that it’s likely to rain today, that Christmas is going to be on a Thursday this year, etc. These are beliefs that—for us, at least—require inferential support. *Any* theory of justification that doesn’t explicitly single out a class of beliefs as requiring doxastic/inferential justification is in danger of refutation from such examples. (2011b: 477)

Here is Lyons’s recipe for these examples:

> RECIPE ONE: First, take any belief that is clearly only justified inferentially *for us* if justified at all; take any “non-basic” belief you please. Second, stipulate that one of us, by some fluke, mutation, or even benevolent intervention, acquires a reliable process that causes a belief like that without any inferential support from other justified beliefs or any other “evidential” support. That belief so formed, clearly, is not *prima facie* justified.

The two most famous examples are Bonjour’s Norman (Bonjour 1980) and Lehrer's Truetemp (Lehrer 1990). Here's a detailed version of Bonjour's case, with everything the experientialist cares about screened off.

> NORMAN, an otherwise ordinary four-year old boy, just so happens to have a reliable "clairvoyant" belief-forming cognitive system in his head with hidden sensory transducers, due to some bizarre and completely random mutation or neurosurgical prank. This process reliably induces true beliefs about the whereabouts of the American President; his beliefs are true in the actual and in nearby possible worlds. The mutation reliably tracks the President, partly because clairvoyance waves have recently filled our atmosphere (also by cosmic accident), and the President emits signals carried by those waves (again by cosmic accident).
>
> Norman has no meta-beliefs about his possession of this process, nor does he have any meta-beliefs about the reliability of such processes.
>
> Unlike many other belief-forming processes, this one entirely lacks any accompanying conscious sensations, conscious representations, or other "seeming-to-be-true" phenomenology. All the process does is stick true beliefs in Norman's head, without his awareness or acknowledgment. They don't even seem to come to him from out of the blue; he's got no clue that he's formed such a belief or why. It's as if they've been there all along.
>
> These beliefs play no significant role in his life or overall mental economy. He receives no feedback of any sort or in any way that's he's right; these beliefs are otherwise entirely idle. He does nothing with the information; it serves no intellectual or practical end.
>
> Even so, the belief is accessible, like stored beliefs in memory, just not its source or basis. So if you were to ask Norman where he believed the President was, he could tell you. If you then were to ask him why he believed it, he may confess he had no idea; we often forget the sources of our beliefs. "I don't remember." Or he might confabulate reasons that, in fact, have nothing to do with why he believes what he does: a common occurrence. "My uncle must have told me." (Nisbett & Wilson 1977)

In this example everything the run-of-the-mill reliabilist wants is present, but everything the experientialist wants is not. And, as anyone familiar with the literature knows, nearly everyone believes Norman's beliefs, despite reliably formed, are not justified, including leading reliabilists. The reliability of the process *may* take Norman's beliefs close to being *knowledge*. But the reliability does not *ipso facto justify* his beliefs. Reliabilists then face a big problem: *in situ*, matter of fact reliability does not seem sufficient for justified belief. Norman's beliefs, *if they are to be justified*, require inferential support from other justified beliefs. *De facto, in situ* reliability is not sufficient to *prima facie* justify his beliefs.

Lyons agrees. Norman's beliefs, though reliably formed, are "not *prima facie* justified, and thus the case is a counterexample" to simple reliabilism (Lyons 2009: 114, 118). The qualification "prima facie" matters, for as Lyons rightly observes, Norman's case is not simply one of *prima facie* justification defeated by other things he believes or should believe. Rather it's the complete lack of *prima facie* justification in the first place that's at issue. Your theory of defeaters won't solve the problem posed by clairvoyant powers cases (2009: 123-5).[2)]

[2)] I realize the case is not always understood as a problem "internal" to reliabilism. In part that is because experientialists tend to add that Norman's belief comes to him "out of the blue" like a sudden attack of heartburn or some strange premonition, where it would be natural for Norman to feel uncertain and to wonder why he believes as he does. Experientialists tend to emphasize that from Norman's conscious perspective, it seems entirely accidental to him that he believes the President is in New York. This description, however, tends to assimilate Norman's case to Bonjour's other cases where the subject's belief is defeated by other things the subject believes or should believe, given other elements within his or her perspective. Reliabilists then tend to be unmoved by the case, as they feel satisfied with their account of defeaters. So in order to make clear that we are focusing on reliabilism as theory of *prima facie*

Lyons concludes that Bonjour's Norman poses a real problem for reliabilism, for it's incredibly easy to construct counterexamples by this recipe (2009: 122, 135). Pick your inferential belief, pick your mutation, and you're off to the races. Lyons believes that some beliefs require inferential justification and some do not. For those that do not, the reliability of the process is sufficient for *prima facie* justification. For those that do, justified supporting beliefs must be called to muster; otherwise the belief is not justified, even if reliably formed. Reliabilists have been "unacceptably lax" in their "treatment" of inferential justification.

This led Lyons to propose and defend "Inferential Reliabilism" against "Simple Reliabilism." According to Simple Reliabilism reliability is sufficient for justifiedness:

> **If** S's belief that p results from results from an *in situ* reliable cognitive process, **then** S's belief that P is *prima facie* justified.

According to Simple Reliabilism, Norman's beliefs about the President should be *prima facie* justified, when clearly they are not.

Lyons's alternative has a number of conditions. Here is the first, designed to exclude Norman, for Norman's belief does not satisfy the antecedent, for reliability alone is not sufficient for justifiedness.

> (1) **If** S's belief that p is the result of the non-inferential operation of a primal system, and the operation of the process is reliable *in situ*, **then** the belief that p is *prima facie* justified.

justification, I have explicitly eliminated any such feeling from Norman's perspective that something is not quite right.

To understand this, you need to know what a *primal system* is. Lyons will construct a theory of primal systems and basic beliefs (hence the title of his book, *Perception and Basic Beliefs*) that entails Norman's belief is not the result of a primal system. Non-inferentially and reliably formed, the belief is nevertheless not "basic" for not formed on the basis of a primal system. Non-basic beliefs then make up the class of beliefs that need inferential justification if they are to be justified at all. Norman's beliefs lack inferential justification, and for that reason isn't justified, even if reliably formed. A correct theory of when a belief needs inferential justification then "solves" the clairvoyance problem for reliabilism. Hence the title of the theory: "Inferential Reliabilism."

II. Primal Systems and Basic Beliefs

What, then, is Lyons's theory of basic beliefs? Lyons first develops a theory of perceptual belief, where a perceptual belief is simply the output of a perceptual system, where a perceptual system is a cognitive system such that:

> (a) Its lowest level inputs are transducers across sense organs.
> (b) None of the inputs to any of its subsystems is under the voluntary control of the larger organism.
> (c) It is "inferentially opaque" (i.e. its doxastic outputs are cognitively spontaneous; they are not the result of an introspective train of reasoning from earlier beliefs; the only introspectively accessible inter-level representations produced by the system are nondoxastic; none of its inter-level representations are conscious beliefs). (2009: 95, 136)
> (d) It has a "normal" etiology; i.e. it results from an interplay of learning and innate constraints.

Lyons then generalizes from perceptual systems to “primal” systems:

> The theory…generalizes... I call a system that satisfies (c) and (d) above a “primal system”, as the term is suggestive of both the ontogeny and the opacity of the system. Conditions (a) and (b) are distinctive of perceptual systems and are not required of all basic-belief-producing systems. (2011a: 445)

Conditions (c) and (d) are the defining features of “primal” cognitive systems. Condition (c) allows that sometimes the (opaque) inputs to a belief are other beliefs. When that happens, the operation of the (opaque) system is (partly) inferential. But when this does not occur, the system operates non-inferentially.

With primal systems in hand, Lyons then defines basic belief:

> A belief B is basic for S at t iff B is the output of one of S’s primal cognitive systems that (i) is inferentially opaque, (ii) has resulted from learning and innate constraints, and (iii) does not base B on any doxastic inputs at t. (2009: 144)

You might have thought (I know I did) that the distinction between basic and. non-basic beliefs was entirely architectural: basic beliefs are not inferentially based on other beliefs, non-basic beliefs are. But Lyons thinks you’d be wrong, or only half right. Origins matter too: basic beliefs arise from “primal” cognitive systems, systems that are either innate or learned or a bit of both. A non-inferential belief caused by a system with the wrong origins—the wrong etiology—isn’t a basic belief, no matter the architecture of the system (2009: 126). So “basicality” for beliefs has (at least) two dimensions: architecture (is the belief inferential or non-inferential?) and origins (is the system innate or appropriately acquired?).

According to Inferential Reliabilism, Norman's clairvoyance beliefs are basic prima facie justified *if* they meet four conditions: opacity, etiology, non-inferentiality, and reliability. Norman's beliefs meet opacity, non-inferentiality and reliability (for the system is opaque, it operates non-inferentially, and it is reliable), but not etiology (origins), for the system is brand spanking new. Norman's clairvoyance *beliefs* are thus not *basic* for the source is not *primal*. They thus fail to be non-inferentially justified, even if non-inferentially and reliably formed. Reliability alone is not sufficient for beliefs about the President, or four-dimensionalism or the age of the earth, for those beliefs are *non-basic*. "For a cognizer built like us, there are simply some propositions that can't be justified without evidential support from other beliefs. These beliefs are non-basic for us" (Lyons 2009: 122-4, 144). Simply causing them reliably is not sufficient for *prima facie* justification. Lyons has spelled out a class of beliefs (the non-basic beliefs) as requiring doxastic/inferential justification by spelling out the class of basic beliefs justification. Basic beliefs do not require inferential support from other beliefs; non-basic beliefs do.

Perceptual beliefs (for us) are basic (for our perceptual beliefs have the right origins and so do not need inferential support from other justified beliefs); clairvoyant beliefs (for us) are not (for they have the wrong origins and so would need support from other justified beliefs; a reliable mutation would not be enough). The clairvoyance problem is really about "basicality" and not about reliability. "What looked to be objections to the claim that⋯reliability is sufficient for the *prima facie* justification of some beliefs begin to look like objections to the claim that some particular belief is basic" (Lyons 2009: 166). The problem with Norman is that he is non-inferentially forming beliefs from non-primal systems that, for us, must be formed inferentially to be justified. Counterexample diffused. Reliabilism is no longer in danger of refutation from clairvoyance examples.

III. Basic Inference

Not just yet. Notice how Lyons's solution works: Inferential Reliabilism blocks the recipe of creating justified *basic* beliefs on Simple Reliabilism by restricting what counts as a *basic* belief. Since Norman's belief is not basic (the system is not a primal system for it has the wrong origins), it's not a problem for Inferential Reliabilism. Mutations needn't pose a problem, for mutations don't create primal systems. A "system that just came into being overnight would fail to satisfy the etiological constraint" (2009: 136-7). Hence they don't create basic beliefs, even if they create non-inferentially formed, reliably true beliefs.

But what if a belief is *inferentially* based on a justified belief (and so apparently has all the inferential support it needs) but results, in part, from a bizarre mutation? Won't that cause the same problem all over again? Instead of imagining our protagonist forming beliefs *non*-inferentially as the result of a mutation, imagine she forms them *inferentially* as the result of a mutation.

Here's a recipe for such cases.

> SECOND RECIPE. First, imagine a justified basic belief (or a set of such beliefs). Second, imagine a belief that, if justified for us, is only justified inferentially (or a set of such beliefs). Third, imagine a mutation that produces a conditionally reliable process that takes the former basic justified belief (or set) as input and reliably produces the latter belief (or set) as

> output, such that the output is conditionally reliable on the input.

If mutations cause a problem for non-inferentially formed beliefs, they should cause the very same problem for inferentially formed beliefs.

Here's a concrete example.

> NORMALA is a four year-old girl. She reliably forms perceptual beliefs about the shape of surfaces (as we all do). Imagine she forms the perceptual belief *that surface is round*. Then, according to Inferential Reliabilism, that belief is *prima facie* justified.
>
> Then imagine, due to a strange and bizarre mutation, this *prima facie* justified basic belief causes, without Normala's awareness or any accompanying phenomenology or experiential evidence, a reliably true belief about the whereabouts of the American President (or any hard-won item of science or philosophy, or any other belief that, for us at least, requires inferential support). In other words, imagine the mutation causes a reliable *inferential* cognitive system that takes justified beliefs as inputs and produces any belief you please—though reliably true—as output. The mutation, in this case, thus forms a reliable *inferential* belief from a *justified* basic belief.
>
> *You* may resist calling this belief inferentially formed; *Lyons* would not. You may think Normala should "appreciate" the connection between the premise and the conclusion for it to be an inference; Lyons would not.
>
> Normala has no meta-beliefs about her possession or use of this process, nor does she have any meta-beliefs about the reliability of such processes (cp. Lyons 2009: 138-9).
>
> Unlike many other inferential belief-forming processes, this one entirely lacks any accompanying conscious sensations, conscious representations, or other "seeming-to-be-true" phenomenology. The process can be entirely unconscious (Lyons 2009: 139). All the process does is inferentially generate beliefs in Normala's head on the basis of non-inferentially justified basic beliefs, without her awareness or acknowledgment.
>
> These beliefs play no significant role in her life or overall mental

> economy. She receives no feedback of any sort or in any way that she's right; these beliefs are otherwise entirely idle.

The resulting *inferentially* formed belief (for Lyons) seems no more justified than Norman's *non*-inferentially formed belief. Is this case a problem for Lyons? No it isn't, and it is important to see why.

To ward off Normala cases, Lyons requiresthe *right origins* for *inferentially* formed beliefs. Though Lyons didn't consider Normala type cases, he would insist that Normala's inferential transition is not a *basic* transition. Normala's belief, Lyons would say, does not result from a *basic inference*. A "basic inference is one that results from the inferential operation of a primal system (a non-basic inference is any other inference)" (Lyons 2009: 171). Since primal systems are innate or learned, but Normala's inferential process that transitions from justified premises (inputs) to reliably true conclusions (outputs) results from a random mutation, her inferential process is not a primal system. Her inference is then not a *basic inference*. Hence, for Lyons, though it is an inference, it is not a basic inference, and thus the resulting belief is not justified.

Lyons adds the following to Inferential Reliabilism.

> (2) If S's belief that p is the result of the inferential operation of a primal system Σ, where (i) Σ bases the belief that p on the input beliefs that q1,… qn, (ii) the process resulting in the belief is conditionally reliable, and (iii) S is *prima facie* justified in each of q1,…qn, then the belief is *prima facie* justified. (2009: 177)

Since Normala's system isn't a primal system, it does not satisfy the antecedent of (2). Normala's inferential, reliably true belief is then not justified on Lyons's theory, for it is not based on a primal inferential system, just as Norman's non-inferential, reliably true belief is not justified either for it is not based on a primal non-inferential system. Pretty nifty.

IV. Architecture vs. Origins

So far so good. Norman's belief is not basic (for the source has the wrong origins) and Normala's inference is not basic (for the source has the wrong origins). There's a lesson here to be learned. Lyons laments reliabilists unacceptable laxity about *inferential* justification. But the problem he's addressing, I believe, isn't really laxity about "architecture." It's really laxity about *origins*. What "mutation" and "benevolent manipulation" cases show is that origins matter. Clairvoyance cases are *bad* origins cases. It is not simply that reliabilists have been unduly lax about inferential versus non-inferential justification (even if they have), it's rather that they've been unduly lax about origins. The wrong origins can create an unconditionally reliable non-inferential mechaninism or a conditionally reliable inferential mechanism; either way it has the wrong origins.

We can make this point by constructing a clairvoyance counterexample that *doesn't* start with a belief that is intuitively only inferentially justified for us. That will show the problem is about origins, not architecture. Here's a third recipe.

> THIRD RECIPE. First, take a belief (or an analogue of such a belief) that would typically be *non*-inferentially justified for us. Second, reliably cause that belief (or the analogue) through a mutation, where the agent also does not in fact receive any other relevant feedback from other sources or through

> behavior that the belief is true; it's an "isolated" perceptual belief. Voila, another clairvoyance counterexample to Simple Reliabilism.

Usually we form perceptual beliefs on the basis of conscious visual perceptions. But Lyons believes consciousness is inessential for justification: blindsighters—even zombies without any conscious experiences—have basic justified beliefs too. (Hence the subtitle of his book.) So imagine a belief that's nearly a perceptual belief, that the agent doesn't act on or receive any feedback for, and then give it the wrong origins.

> NORBERT is a four year-old boy. Norbert sometimes non-inferentially forms the belief *that surface is round*, formed without any accompanying conscious experience or awareness. He's partially blindsighted.
>
> Imagine furthermore that, in this particular case, it is not caused by a reliable innate or learned perceptual belief-forming capacity, but instead is caused by a reliable cognitive system with hidden and unnoticeable sensory transducers that results from a strange and bizarre mutation.
>
> Norbert enjoys no other collateral epistemic support for this belief. Norbert has no meta-beliefs about his possession of this process, nor does he have any meta-beliefs about the reliability of such processes. All the process does is stick these beliefs in his head, without his awareness or acknowledgment. They don't even seem to come to him from out of the blue; he's got no clue that he's formed such a belief or why. It's as if they've been there all along.
>
> Unlike typical perceptual beliefs, this belief plays no significant role in his life or overall mental economy. He receives no feedback of any sort or in any way that's he's right; the belief is otherwise entirely idle.

Though Norbert's belief has the same kind of content as a typical perceptual belief, it is not formed by a (strictly speaking) "primal" system (and so, for Lyons, not strictly speaking a perceptual belief). It is, however, non-inferentially and reliably formed.

Lyons would not call this belief justified, for it has the wrong etiology. Norbert is in the same boat as Norman and Normala. So take any belief that, typically, is non-inferentially justified for us (or an analogue), but then give it the wrong origins, and make sure not to give it any other epistemic support that Lyons has antecedently screened off as irrelevant. We have the same problem all over again. You don't have to start with a belief that is only inferentially justified for us to cause havoc for reliability theories of justification with bizarre mutations.[3)]

It should be clear by now that the clairvoyance problem is fundamentally about *origins*, not about *architecture*. The issue isn't simply about the structure of the building; it's more importantly about why the building is there in the first place. Both Norman's and Norbert's beliefs were *non*-inferentially formed with the *wrong* origin and *not* justified, according to Lyons. Normala's belief was *inferentially* formed with the *wrong* origin and also *not* justified, according to Lyons. Inferential or not, wrong origins excludes justification. But then the problem with reliability theories of justification is not simply that they've been unduly lax in their treatment of *inferential* justification. The deeper problem is that they've been unduly lax in their treatment of *origins*. A theory of justification that doesn't explicitly single out the right origins is in danger of refutation from such examples.

[3)] Lyons thinks Bonjour and Lehrer stacked the deck in their famous examples by starting with beliefs that are clearly only inferentially justified for us (no wonder Bonjour and Lehrer think those beliefs require inferential support from meta-beliefs). Lyons then challenges Bonjour and Lehrer to concoct a case that challenges reliabilism without beliefs that are clearly non-inferentially justified for us. Now of course Bonjour and Lehrer won't take up that challenge, for they deny the very existence of non-inferentially justified (empirical) beliefs. But since we're not so committed, we can concoct such a case. Norbert should do the trick.

V. The Relativity of Origins

The preceding should be enough to establish my first thesis; the real issue isn't architecture but origins. More can be said to make this point stick. For Lyons allows psychological duplicates with the same architecture to differ in epistemic status because of different origins. It shall prove worthwhile to spell this out.

Recall Norman. Perception is primal for Norman, but clairvoyance is not. Norman enjoys perception innately. Norman's perceptual system goes through normal stages of development. Norman also learns new perceptual categories and expert perceptual categorization through learning. Norman's perceptual systems satisfy Lyons's etiological constraint on primal systems. Norman does not enjoy clairvoyance innately. Nor did he acquire it through normal stages of development from other innate systems. Nor did he develop clairvoyance through any learning mechanism. Norman acquired clairvoyance through some random mutation, perhaps by stepping in radioactive waste. Clairvoyance just popped into his head one day, without his knowledge or acknowledgement.

According to Lyons, perceptual beliefs (for us) are basic, but "clairvoyant beliefs (for us) are not. Perceptual beliefs are the outputs of a [primal] system; clairvoyant beliefs are not" (2009: 121). Given the way we are built, some beliefs are basic and some are not. "For a cognizer built like us, there are simply some propositions that can't be justified without evidential support from other beliefs. These beliefs are non-basic for us" (2009: 122-4, 144).

However, the exact same clairvoyant system Norman acquires may be innate in some other possible individual or species of individuals. Lyons gives the following example of such an individual:

> NYRMOON is a four year-old boy. He's a member of an alien but human-like species, living in a different environment in a different possible world. Nyrmoon's species have clairvoyance as a normal, reliable cognitive capacity, "which develops in much the same way as vision does for humans. Members of Nyrmoon's species have specialized organs that are receptive to the highly attenuated energy signals from distant events. (Lyons 2009: 119; Sosa 1980; Goldman 1988)

Clairvoyance is primal for Nyrmoon, just like perception for Norman. Nyrmoon enjoys clairvoyance innately. Nyrmoon's clairvoyance system goes through normal stages of development. Nyrmoon also learns new clairvoyant categories and expert clairvoyant categorization through learning. Nyrmoon's clairvoyant systems satisfy Lyons's etiological constraint on primal systems.

Nyrmoon's entire species works this way. They all have it innately, and they all go through normal stages of development (Lyons 2009: 144, 164). It reliably tracks features of their environment, and they rely on this capacity to navigate and flourish in their natural habitats. Clairvoyance is "basic" for Nyrmoon and his species; clairvoyance produces basic beliefs *for them* (Sosa 1980, Goldman 1988).

Can Nyrmoon form justified "clairvoyance" beliefs? Lyons believes he can. I concur. If we have justified basic perceptual beliefs, then surely Nyrmoon has justified basic clairvoyance beliefs. This non-actual, reliable belief-forming process is just as good as ordinary human perception. Like human perception, it has the right origins. Clairvoyance is "basic" for Nyrmoon's species but not for ours. Clairvoyance satisfies the etiological constraint *for them* but not *for us*. Clairvoyance produces basic beliefs *for* *them* but not *for us*.

We can even imagine that Norman and Nyrmoon are molecule-for-molecule duplicates. The point is the same: Nyrmoon's beliefs are justified; Norman's are not. Nyrmoon's species, like ours, relies on vision for what they can see. Unlike us, they also use clairvoyance for what they can't. Clairvoyance for them has the *right* origins; clairvoyance for us has the *wrong* origins. Same "architecture" different origins. What really matters to the reliabilist program is getting the origins just right. The preceding sections should be more than enough to establish my first thesis: it's not laxity about architecture that matters, but laxity about origins. What a reliability theory needs to rule out strange origin cases is a good theory of good origins, not simply more attention to the contrast between inferential and non-inferential architecture.

VI. Why Origins Might Matter

But if origins are what really matters, why do they matter? What explains why a belief (either non-inferential or inferential) must be based on a *primal* system? I now turn to my second thesis. I will argue that, unfortunately, Lyons never really explains why the origins he selects matter to turning reliably formed but unjustified beliefs into reliably formed justified beliefs; Lyons doesn't say why innate or learned reliable non-inferential systems or why innate or learned inferential systems are the right kind of systems. His argument for origins is entirely case-based: if mutations make for bad origins, then non-mutations must make for good origins. That's all we get from Lyons. But if innateness and learning matters, are there reasons for thinking they do?

Here is a fundamental feature of many innate systems that he might have found relevant for thinking that innateness matters for justified beliefs.

Many traits are *adaptive* in the sense that they are useful to the organism. One clear way to be biologically useful is to contribute to relative fitness; you are more likely to survive and reproduce. When an organism has an adaptive trait, the organism can either have it *because* it is adaptive or for some other reason. When an organism has a trait *because* it is adaptive, the trait is an *adaptation*. Organisms have many of their adaptive traits because they are adaptive; many adaptive traits are adaptations. An *adaptive explanation* explains why a trait exists *because* it is adaptive.[4]

Evolution by natural selection is the only non-magical explanation of the existence of innate, functionally complex adaptive traits, especially those that have evolved through convergent evolution (Dawkins 1986). I

[4] There is a tendency among philosophers skeptical of evolution and evolutionary explanations to cry "Gould!" in the hopes of screening off evolutionary explanations, even in "naturalized" epistemology (e.g. Lyons 2009: 128). Gould spent a good deal of his career arguing against strong adaptationism in biology, the doctrine that all traits—especially obviously adaptive traits—arose due to natural selection for their currently adaptive effect. Gould argued instead that many adaptive traits are "exaptations" and that many other traits are not adaptive at all. He called some of these latter traits "spandrels." Gould's influence here was salutary; adaptationists sometimes go too far. But it should not distract from the well-entrenched fact that many traits are adaptations—especially the functionally complex and those that arise through convergent evolution—where eyes through the animal kingdom are a paradigm case. (Gould, for one, would never, ever deny that eyes are adaptations. Gould is not an anti-evolutionary lunatic.) Furthermore, exaptations are often just as much adaptations as adaptations; they are simply traits that originally arose for some other adaptive purpose that were later modified and adapted to some other, or some additional, purpose. For discussion and references, see Sterelny and Griffiths 1997 and David Buss, et. al., 1998.

confidently assert that every innate complex cognitive system we've got that's reliable and clearly confers justification—especially perception—resulted from evolution by natural selection. Bracketing magical thinking, Nyrmoon's clairvoyance (like Norman's perception) would have resulted from natural selection too. Natural (directional and maintenance) selection is the best explanation for the origin and persistence of functionally complex adaptive traits in a population, especially those that arise across the animal kingdom through convergent evolution.

Natural selection works by taking variants of traits and selecting among the variants because of their consequences. The variants with the best relative consequences are preserved. Natural selection is then a feedback mechanism; it takes relative consequences of ancestor traits as input and produces the descendent traits with those same consequences as output. Perception, memory, and reasoning are all adaptations (they're certainly not spandrels; that would be absurd). They have all resulted from evolution by natural selection. Perception reliably induces true beliefs, and by doing so it contributes to relative fitness, and by doing so it contributes to its persistence in the human population. Innate reliable psychological processes are then acquired, in part, because they conduce good consequences by being reliable, where those good consequences enter into a feedback loop that explain why the processes are selected and retained.[5]

Perception is not merely "truth" adaptive in the sense that it reliably causes and sustains true beliefs. Perception is an *adaptation*. Perception

[5] Natural selection certainly explains the change in frequency—and so the persistence—of a trait in a population. Without natural selection there wouldn't be visual systems across the animal kingdom. Does it also explain, at least in part, why an *individual* organism has a *token* of that trait? Cummins (1975) and Sober (1984) have argued it does not. Nanay (2005) convincingly responds.

exists, in part, *because* perception is useful. Perceptual systems exist, and we have them, partly because they confer benefits on us, partly by reliably causing and sustaining true beliefs. Getting things reliably right then partly *explains why* we have these processes; feedback matters.

Biological adaptations exist, in part, because they are adaptive; they are a good fit between an organism and its environment. Adaptations are those aspects of the morphology, physiology, and behavior of organisms that are adaptive solutions to problems posed by the environment, adaptive solutions that arise and persist because of an explanatory history of evolution by natural selection, that arise and persist partly because of a good fit.[6)]

So Lyons may think innateness matters because those reliable processes that are innate in us are there because they have entered feedback loops that contribute to their continued existence. Lyons may think innateness matters because innate cognitive processes are not merely truth-adaptive but genuine adaptations. They are not just good fits between mind and world, but they arise and persist *because* they are good fits.

Here is a fundamental feature of many learned traits and behaviors that he might have found relevant for thinking that learning matters for justified beliefs.

Psychologists and ethologists see learning, like evolution by natural

[6)] I discuss objections to the claim that evolution would select for reliably true psychological processes in my 2014a. Is Lehrer's tempucomp in Truetemp's head an adaptation? No. Though the device has an assigned function from the designer, and though it may reliably fulfill its function, it has not entered into a feedback mechanism that explains why it persists in terms of its benefits to Truetemp. In fact the case is described so that it doesn't. Intentionally assigned function is one thing, adaptation is another. Cp. Lyons 2009: 128. I discuss the Truetemp case in more detail in my 'Proper Functionalism and the Proper Theory of Functions.'

selection, as a process that produces adaptations to the environment: useful traits that persist because they are useful. Learning is a feedback mechanism that takes input from the environment and produces adaptive solutions as output; learning is the adaptive modification of behavior based on experience (Lorenz 1966, Alcock 2009: 97-98). Trial-and-error learning (either conscious or unconscious) is the paradigm for learning new skills or acquiring new "systems."[7] It is a trite observation in psychology textbooks that trial-and-error learning, even if very fast, resembles evolution by natural selection. For it involves variants in behavior, consequences of that behavior, and then modification of future behavior (and so selection among variants) on the basis of the consequences of that behavior. So if you are learning or acquiring a new belief-forming process, you'll first make mistakes, and then get feedback on the basis of which you'll modify your behavior or belief-forming processes until you settle on the one you have thereby "learned" to use. Learned reliable processes are then learned, in part, because they conduce good consequences by being reliable, where those good consequences enter into a feedback loop that explain why the processes are modified and then retained. Learning produces a good (an adaptive) fit. Learning mechanisms adapt the organism to its environment. Learning mechanisms produce adaptations, good fits that exist partly because they are good fits.

Learning mechanisms are themselves innate adaptations, innate mechanisms for further adapting the organism to its environment. "If learning is an adaptive improvement, there has to be, in Lorenz's phrase,

[7] "Perceptual systems develop through the interaction of genetics and environmental factors, a combination of learning and innateness. Experience fine-tunes discriminatory abilities⋯Learning can also result in [whole new systems]" (Lyons 2009: 92, 95).

an innate teaching mechanism, or "innate schoolmarm.""(Dawkins 2010: 361). Nature encodes learning mechanisms when the environment is sufficiently unpredictable to simply encode solutions to environmental problems. If an organism needs to adapt to its changing environment in its lifetime, nature builds in learning mechanisms.[8]

So Lyons may think innateness and learning matters because those reliable processes (either innate or learned) are there because they have entered feedback loops that contribute to their continued existence; we have these capacities because they are reliable, and that's what (at least in part) explains why *these* (innate or learned) as opposed to *those* (mutations or surgical interventions) origins matter. The reliable processes that confer justification exist partly because they reliably produce true beliefs; reliable processes that do not confer justification do not exist or persist (partly) *because* they are reliable. Being reliable explains why they exist; being reliable is then *ipso facto* a non-accidental, explanatory property of the system. Inferential and non-inferential cognitive systems confer justification when they are adaptations *for* reliably causing and sustaining true beliefs, for then the reliability of the system is a non-accidental feature of the system. Good fits are not just adaptive fits; good fits are adaptations. These systems are non-accidentally reliable.

Norman's clairvoyance is just an accident, even if a good fit. He doesn't have it because it is adaptive; its reliability does not explain why he has it. Nyrmoon's clairvoyance, on the other hand, isn't just an accident. He has it (partly) because it is adaptive; its reliability does explain why he has it. Norman's reliability is an *accidental*, *non*-explanatory fact. Nyrmoon's reliability is a *non*-accidental,

[8] I discuss learning, natural selection and feedback mechanisms at greater length in my 2014b and my in preparation.

explanatory fact. Nyrmoon’s process is an adaptation; Norman’s is not. Norman’s clairvoyance case shows accidental reliability is not sufficient for justification. Maybe that’s why origins matter. Or maybe that is at least a step along the way to explaining why origins matter.

VII. Hopeful Monsters

But this is not at all why Lyons thinks. For Lyons doesn’t think innateness or learning matters *because* innate are learned traits are mostly adaptations (and so explained by their beneficial effects through feedback mechanisms). Rather he thinks innateness matters because….

Actually, he doesn’t say. He gives us no hints at all. He gives us no direction for constructing an explanation for why innateness, for example, might really matter. In fact, it seems, any “innateness” will do, even “innate” traits *without an explanation*. Consider the following apparent counterexample to Lyons’s view.

> NORCO is a four year-old boy who has a clairvoyant powers mutation written into his genes. Maybe the mutation occurred during conception, or early in the pregnancy. Or maybe it occurred late in the pregnancy, or just before (even seconds before) birth. Imagine too that the mutation lies dormant for years only comes to fruition later in life, at exactly the same time as Norman’s mutation, affecting Norco exactly the same was as Norman. We can even image that when the mutation comes to life Norman and Norco are molecule-for-molecule duplicates.
>
> Norco’s innate mutation then reliably induces true beliefs about the whereabouts of the American President, exactly like Norman’s non-innate mutation. Like Norman’s, it entirely lacks any accompanying conscious sensations, conscious representations, or other “seeming-to-be-true” phenomenology. They don’t even seem to come to him from out of the blue;

> he's got no clue that he's formed such a belief or why. These beliefs play no role in his life or overall mental economy. He receives no feedback, of any sort or in any way, that's he's right; these beliefs are otherwise entirely idle.
>
> No selection of learning of any kind explains why Norco has this process. It is reliable, but its being reliable is explanatorily irrelevant to its existence or persistence. Its reliability does not explain, in any way, why he has this process.

Now it seems to me—I'm speaking from the heart now—that if Norman is in trouble, then so is Norco. If the wrong origin—a lucky mutation—rules out Norman's reliably true beliefs as justified, then the wrong origin—a lucky mutation—rules out Norco's beliefs as justified. We have even imagined they become duplicates. Norco's clairvoyance is written into his genes before birth but only emerges later. Norman's identical clairvoyance only gets into his genes later in life. What's the difference?

If being reliable doesn't enter into a feedback loop that explains why the individual has the system, then the possession of the system isn't explained by the system's being reliable, and then the reliability is a non-explanatory, accidental feature of the system. If Norman doesn't have what it takes for justified beliefs about the location of the leader of the free world, then neither does his doppelganger Norco. If the reliabilist is worried about mutations, "accidental" causes of psychological systems—even if they reliably induce true beliefs—then *when* they occur should not matter.

Norco strikes me as a clear counterexample to Lyons's account of justified basic beliefs. On Lyons's theory, Norman's beliefs are not basic for not the result of a primal system, for Norman's process isn't innate.

On the other hand, Norco's beliefs are basic on Lyons's theory, for Norco's process is innate. So on Lyons's theory, Norman and Norco do not stand or fall together, for Norman was mutated after birth, and Norco was mutated before. *Innate* origins are not the same as the *right* origins.

Norco's mutation is what a biologist might call a "hopeful monster." In nature, new traits often arise very slowly through a series of cumulative micro-mutations. But once in a while a macro-mutation arises, where a new trait emerges in a single step. In almost every case these macro-mutations are harmful; think of extreme birth defects. But in principle, and sometimes in practice, these macro-mutations are beneficial to the organism, hence the name "hopeful monster." Norco's mutation is then "truth-adaptive" in the sense that it reliably induces true beliefs, even though, *per hypothesis*, Norco doesn't benefit in any other way from the mutation (he receives no feedback and the beliefs are otherwise idle), and so his "truth-adaptive" mutation isn't, in any sense, an adaptation.

When confronted with a case like this on another occasion, Lyons confesses that he has the "wrong" intuition: "the mere fact" that an individual like Norco is "the first of his line" (and, for all that, he may be the last), "doesn't affect my intuition that he's justified, so long as we conceive his clairvoyance module to be a primal system in my sense" (Lyons 2011b: 486).

Lyons thus doesn't think it matters *why* the process is innate. All that matters, for Lyons, is that it the system *is* innate, a part of the individual's "basic" package. Any kind of innateness, it seems, will do the trick. It is perfectly OK with Lyons if the process is accidentally reliable, a cosmic accident, provided only that it is "written in to the genes" *before* birth.

VIII. Origins Emasculated

Lyons not only thinks Norco fits the bill, but he also "has the intuition that Swampman [a molecule-for-molecule duplicate of you or me, created by random accident when a bolt of lightning hits a log in a swamp] has justified beliefs, even though he has no phylogenetic history" (Lyons 2011b: 486). And Swampman at creation doesn't have any ontogenetic history either; he's yet to learn a single thing. He certainly hasn't modified any of his innate systems through development. He has no phylogentic *or* ontogenetic history, for he has no history at all. No consequences—no feedback—explain anything about him at all. He's a cosmic accident, a mystery of mysteries.

Even so, Lyons thinks Swampman's capacities are "innate." To "say that a trait is innate is—very roughly—to say that its presence in an entity was more or less determined by the initial state of that entity…The Swampman's initial state is the state he is in when he comes into existence. What systems he has at that time, he has innately, so the Swampman is guaranteed to have systems that satisfy the etiological constraint if he has any systems at all" (Lyons 2009: 147). Swampman meets the etiological constraint, Lyons claims, and so he has justified beliefs on Lyons's theory.

Swampman? Really?

Lyons argued for an etiological constraint, but then when the chips are down the only etiology he cares about is the *initial state* of the organism, never mind the *explanation* for the initial state. Lyons called in Swampman to sheer off phylogenetic history as irrelevant to justification. Instead he says he only favors a "narrow" etiological constraint that "is only concerned with the ontogenetic history of the organism" (Lyons

2011b: 486).

But Swampman is more corrosive than that, for Swampman sheers off learning history as well. For Swampman, like Norman's clairvoyant power, just popped into existence. Learning isn't essential, for Swampman hasn't had a chance to learn anything or to develop in anyway; he's fully formed, up and running. Evolution by natural selection doesn't matter, for Swampman has no history, one in a zillion, possibly sterile and so the first and last of his kind. Learning plays no role, for he hasn't learned a single thing, and a bolt of lightning might kill him off any day.[9] Feedback then plays absolutely no role whatsoever, as far as I can tell, in Lyons's theory. But if feedback plays absolutely no role whatsoever, then why care so much about origins, especially when two origins we know a good deal about—evolution and learning—are feedback mechanisms on beneficial effects?

Swampman might as well be a molecule for molecule duplicate of Norman. Suddenly Swampman has justified clairvoyance beliefs but Norman doesn't? Suddenly Swampman has justified beliefs because he has *no* history (so *ipso facto* everything about him is innate) but Norman *lacks* justified clairvoyance beliefs because he does have a history?

Human perception exists because of feedback. We've got eyes because having eyes made a huge difference in our evolution. We develop new perceptual categories through perceptual learning because of feedback. We learn new categories through perception because making finer discriminations helps us achieve our ends. It's not just that human

[9] True, Swampman might interact with his environment and receive feedback from his environment that explains why he continues to use his seemingly psychological capacities. But Lyons has screened that off. Swampman at the second of creation, before any feedback, has reliable powers with, according to Lyons, the right origins.

perception reliably induces true beliefs; we act on those beliefs and receive feedback on their utility, feedback that in turn explains why we continue to possess, use, and modify those capacities. And what is true of us is surely true of Nyrmoon and his conspecifics. They have clairvoyance because it helps them. They've got clairvoyant powers because of feedback.

I object to the way Lyons emasculates origins to accommodate Swampman. If you are going to reject certain obviously important features of perception in humans and clairvoyance in alien species as irrelevant, then you can't build a theory around cases that involve those features without first screening those features out. After all, part of what makes Norman's case of "bad" clairvoyance so compelling is that he's received no feedback whatsoever, whereas Nyrmoon has. You can't surreptitiously rely on well-established and explanatorily salient features of perception and "good" clairvoyance and then dismiss them as irrelevant when Swampman knocks on your door.

In principle the *intuition* that Swampman has justified beliefs is not so hard to "accommodate." Since we've stipulated that Swampman is a molecule for molecule duplicate of an ordinary human being, and we all know ordinary human beings have justified beliefs, then it will certainly *seem, prior to reflection*, that Swampman has justified beliefs, for Swampman certainly *seems* an awful lot like us. Like fools gold (that looks just like gold to the untrained eye) or twin water (that requires a little chemistry to distinguish from real water), it is easy for Swampman to seem just like you and me without a thorough philosophical examination. But just as we can give up the claim that fools good and twin water are real gold and real water upon further reflection, we can give up the claim that Swampman has justified beliefs, especially at conception, upon further reflection. We may find it “intuitive” that he has justified beliefs, knowing full well that he doesn’t, for he lacks the right etiology. With a good theory of the right origins, you can explain why, though Swampman *seems* to be just like us, in reality he is not. He’s got the *wrong origins*.[10)]

Lyons takes us on a long ride that emphasizes inference and architecture, but the real issues surround origins. But when the chips are down, he says *nothing* about why origins matter. That’s my real beef. An adequate theory needs to say why *these* as opposed to *those* origins make the difference between reliably true justified beliefs and reliably true unjustified beliefs. Why do *these* origins and not *those* produce justified beliefs? Why does being written in to the “initial state” of the organism make all the difference? He doesn’t say.[11)]

[10)] I discuss Swampman at greater length in my 2012, 2014b, and in preparation.

[11)] Thanks to Zach Bachman, Meredith McFadden, Megan Stotts and the referees for the journal that led to a number of improvements.

References

Audi, Robert (1988) *Belief, Justification, and Knowledge*, Wadsworth.

Bach, Kent (1985) “A Rationale for Reliabilism”, *The Monist* 68(2), pp. 246-263.

Bonjour, Laurence (1980) “Externalist Theories of Knowledge”, *Midwest Studies in Philosophy* 5(1), pp 53-74.

Buss, David and T. Shackelford, A. Bleske, J. Wakefield (1998). “Adaptations, Exaptations, and Spandrels”, *American Psychologist* 55, pp. 533-548.

Cohen, Stewart (1983) “Justification and Truth”, *Philosophical Studies* 46(3), pp. 279-295.

Cummins, Robert (1975) “Functional Analysis”, *Journal of Philosophy* 72(20), pp. 741-765.

Dawkins, Richard (2010). “Universal Darwinism” in Mark Bedau and Carol Cleland (eds.), *The Nature of Life: Classical and Contemporary Perspectives From Philosophy and Science*, Cambridge: Cambridge University Press, pp. 360-373.

Foley, Richard (1985). “What’s Wrong with Reliabilism?”, *The Monist* 68(2), pp. 188-202.

Goldman, Alvin (1979). “What is Justified Belief?”, in G. pappas (ed.), *Justification and Knowledge*, Dordrecht: Reidel. Reprinted in A. Goldman, *Liaisons: Philosophy Meets the Cognitive and Social Sciences*, Cambridge, MA: MIT Press (1992).

Goldman, Alvin (1986). *Epistemology and Cognition*, Cambridge, MA: Harvard University Press.

Goldman, Alvin (1992). “Epistemic Folkways and Scientific Epistemology” in Goldman, *Liaisons: Philosophy Meets the*

Cognitive and Social Sciences, Cambridge, MA: MIT Press, pp 155-175.

Goldman, Alvin (1999). "A Priori Warrant and Naturalistic Epistemology" reprinted in A. Goldman, *Pathways to Knowledge: Private and Public*, New York: Oxford University Press (2002).

Graham, Peter J. (2011a). "Psychological Capacity and Positive Epistemic Status" in *The New Intuitionism*, London; New York: Continuum, pp. 128-150.

Graham, Peter J. (2011b). "Perceptual Entitlement and Basic Beliefs", *Philosophical Studies* 153(3), pp. 467-475.

Graham, Peter J. (2012). "Epistemic Entitlement", *Nous* 46(3), pp. 449-482.

Graham, Peter J. (2014a). "The Function of Perception" in A. Fairweather (ed.), *Virtue Epistemology Naturalized*, Synthese Library.

Graham, Peter J. (2014b). "Warrant, Functions, History" in O. Flanagan & A. Fairweather, *Naturalizing Epistemic Virtue*, New York: Cambridge University Press.

Graham, Peter J. (forthcoming). "Against Actual-World Reliabilism" in Miguel-Angel Fernandez (ed.), *Performance Epistemology*, Oxford University Press.

Graham, Peter J. (in preparation). "Proper Functionalism and the Proper Theory of Functions", University of California, Riverside.

Lehrer, Keith (1990). *Theory of Knowledge*, Routledge.

Lorenz, Konrad (1965). *Evolution and Modification of Behavior*, University of Chicago Press.

Lyons, Jack (2009). *Perception and Basic Beliefs: Modules, Zombies, and the Problem of the External World*, Oxford University Press.

Lyons, Jack (2011a). "*Precis* of Perception and Basic Beliefs", *Philosophical Studies* 153, pp. 443-446.

Lyons, Jack (2011b). “Response to Critics”, *Philosophical Studies* 153, pp. 477-488.

Lyons, Jack (2012). “Should Reliabilists Care About Demon-Worlds?”, *Philosophy and Phenomenological Research* 86(1), pp. 1-40.

Nanay, Bence (2005). “Can Cumulative Selection Explain Adaptation?”, *Philosophy of Science* 72, pp.1099-1112.

Nisbett & Timothy Wilson (1977). “Telling More Than We Can Know”, *Psychological Review* 84, pp. 231-59.

Sober, Elliot (1984). *The Nature of Selection*, Cambridge: The MIT Press.

Sosa, Ernest (1980). “The Raft and the Pyramid: Coherence versus Foundations in the Theory of Knowledge”, *Midwest Studies in Philosophy* 5, pp.3-26.

Sterelny, Kim and Paul Griffiths (1999). *Sex and Death: An Introduction to the Philosophy of Biology*, University of Chicago Press.

College of Humanities, Arts and Social Science
University of California, Riverside
peter.graham@ucr.edu

철학적 분석 31 (2014) pp.123-149

The Narrative of Moral Responsibility

Rodrigo Laera

The goal of this paper is to suggest that theoretical thinking with respect to metaphysical determinations or indeterminations is not the appropriate realm for attributing moral responsibility. On the contrary, judgments that attribute moral responsibility (S is responsible for...) depend on the possibility that a rational narrative be built. Agents are capable of forging their future actions, as well as of reflecting upon past actions. With this it will also be shown how we assume control of our behavior because we ignore whether actions are the result of causality or chance. It is claimed that contexts determine the degree of causal demand in narratives that attribute moral responsibility. In order to construct this type of narrative one must focus on a specific link in the causal chain of explanations. If context alone is not demanding enough so as to require that theoretical reflections strive for the ultimate foundation of our actions, then the agent may be considered responsible for his behavior.

Keywords moral responsibility, compatibilism, causal determinism, indeterminism, contextualism

1. Introduction

At first sight, the problem of moral responsibility is based on two conditions:

(a) Actions must be a product of the intentionality of the agent

(b) It has to be known which action caused which consequences

Indeed, from a theoretical point of view it can be argued that for a person to be responsible for an action, said person has to have had some control over the consequence of the action. At the same time it is possible to establish which action caused which event. Both conditions find their axis in the problem of determinism, which consists in whether there is moral responsibility even when human actions are not free - that is to say, the agents not being able to choose between different options for their behavior. Compatibilist theories claim that moral responsibility is possible even though the determinism is true[1]. On the other hand, incompatibilist theories claim that moral responsibility is not compatible with determinism[2].

There are numerous arguments for and against both positions. Unfortunately, there is no room in this paper to even begin to recapitulate the extended and extremely convoluted discussion between compatibilism and incompatibilism. Fortunately, what matters for present purposes are not the details but rather the general character of the discussion: in what contexts can a cause be decisive when attempting to attribute moral responsibility? It is this that will be considered in the following sections.

This paper will argue that moral responsibility is constituted through a narrative of the human actions and not by means of a metaphysic of

[1] E.g., Frankfurt (1971); Wallace (1994) or Fischer (2006)

[2] E.g., Inwagen (1983), Kane (1996), or Peremboon (2001)

alternate possibilities. Furthermore, this paper will show that said narratives have some moral value given that theses narratives could exclude the problem of our actions being metaphysically determined. Thus, the neo-Kantian thesis which states that there is a difference between regarding people from a theoretical perspective than from a practical one can be accepted. In a practical context agents can attribute moral responsibility because they assume that the metaphysical causes do not have a direct and significant impact on our way of being. However, the metaphysical causes are relevant when we examine philosophically demanding contexts on human behavior.

It can be suspected that if a person knew with certainty that our actions are causally determined, the relationship with the moral values would be completely different as to how we nowadays conceive it. However, this suspicion implies a respect for a hierarchy: that the moral sphere depends on the metaphysical sphere in such way that a change in the paradigm of one of them involves a change in the paradigm of the other. Anyhow, it is not necessary to delve further in metaphysical assumptions – especially if it is accepted that context decides how far back we ought to look for causes of an agent's behavior[3].

Waller (2011) has recently claimed that the present scientific understanding of human behavior does not leave space for moral responsibility, and that its abolition is sociologically and psychologically desirable and possible[4]. However, said opinion will be discredited indirectly throughout this paper. On one hand, it shall be argued that compatibilism is possible depending on a context of moral attribution. If

[3] Cfr., Hawthorne (2001), to whom the same action can be both free and un-free (depending on the attributor's context)

[4] See also: Waller (1990)

the context is not demanding enough in such a way that the theoretical reflection goes in search for the final foundation of our actions, then the agent can be responsible for his acts. Thus, the attribution of moral responsibility will depend on rational explanations that refer to actions as being intentional. On the other hand, it shall be explained that the narrative aspect of judgments on moral responsibility helps to understand how it is possible that said responsibility be compatible with naturalism without considering, as Waller claims, that it is based on some miraculous power.

It will be not proposed that moral responsibility depends on whether there is an absence of control, but rather that a person is able to rationally create a narration from which the agent could be capable of both forging future actions and reflecting on past actions.

2. Compatibilism and Incompatibilsm

There exists an ancient skeptic dilemma that claims that subjects cannot be responsible for their actions. The dilemma is the following:

1. If our acts are causally determined, then we are not responsible for them.
2. If our acts are not causally determined, then we are not responsible for them.
3. In consequence, we are not responsible for our acts.

Since the conclusion seems unacceptable, this reasoning has been addressed by accepting the first premise and disambiguating the second premise in such a way that the idea of “cause” does not imply an absolute loss of the freedom to act (e.g. Chisholm, 1966). According to Inwagen (1983), moral responsibility requires that our actions be, at some point, undetermined. Therefore, actions must be voluntary for a person to be responsible for them. Were determinism true in two possible worlds (*M and M'*) in which the same laws of nature are true, then *M* would be exactly equivalent to *M'* in any given *t* moment, and in *any* future moment regarding *t*. For the determinist, acts are conditioned counterfactually, meaning that the laws and the early history of the world are enough to determine the later history[5]. Given that a person is not responsible for what happens before birth, a person will not be responsible for what will happen in the future. Since there is an unavoidable future in regards to moral responsibility there is neither the best nor the worst of the possible worlds.

However, the first premise implies that the agents could choose between alternative acts and that, facing equal acts, those agents can be projected successfully by different possible worlds. This position is represented by the following principle: an agent is morally responsible for what he has done only if he could have done it in any other way; *S* is responsible for an act *A* if he had the option of not doing *A*. In the same way, a person is responsible for not doing a determined act only if he could have done it, since the conditions of moral responsibility are connected to both acts and omissions[6].

It is not a proposition shared by the philosophers that moral responsibility excludes determinism, especially in cases of metaphysical

[5] As states by Lewis (1983): for every historical fact *F* and any starting point in the world, there is a true proposition *H* about the history of *S*, and a true proposition about the laws of nature *L*, in such way that *H* and *L* strictly imply, together, *F*.

[6] In this case, indeterminism is taken as a synonym of libertarianism, meaning that the earlier is based on the latter in a way that, according to both theories, it is necessary some sort of control over our decisions for there to exist morally responsible acts (Cfr., Berofsky, 1995, 2006).

constraint. There exists, then, another strategy to address the aforementioned dilemma, which consists of directly denying the first premise. Consider the cases of the *Frankfurt Style*. A mysterious scientist secretly implants a chip in John's brain so as to supervise and control his actions. Among the things the scientist supervises, there is the taste for the products of a certain brand (*X*). So, if John decides to purchase an item of any other brand (*Y*), the scientist is prepared to intervene by means of sophisticated equipment that he has designed to alter the conduct. On the contrary, if John decides to purchase the items of *X*, then the scientist does not intervene and the equipment keeps on supervising without affecting John's decisions. Now, assume that John decides on his own (as he would do without the intervention of the scientist), to buy an item of the brand *X*. John would be, then, morally responsible for that choice, even if he could not have chosen anything else[7].

Similar cases add up to the compatibilist position[8]. Compatibilists claim that moral responsibility is compatible with determinism, insisting that neither the advance of natural science nor metaphysical perspectives represent a problem for moral responsibility. Starting from this idea, Fischer (2006) has differentiated the regulative control of actions from guidance control[9]. The former encloses a genuine metaphysical access to

[7] So as not to drift away from the objective of the paper, I will leave aside the numerous critics that have aroused this type of cases.

[8] Traditional compatibilism is defined by the conjunction of the following three theses: 1. Free will is essential for moral responsibility; 2. Free will requires that there exist alternative possibilities when carrying out an act; 3. Moral responsibility is compatible with determinism.

[9] Strictly speaking, Fischer is a semi-compatibilist, given the fact that he is a compatibilist in what respects to the relationship between moral responsibility and determinism; and an incompatibilist in what respects to the relationship between determinism and relevant alternatives. In this text, he is taken as a compatibilist since, all in all, semi-compatibilism is simply the affirmation of

alternative possibilities, while the latter is based on the capacity of the agent to be able to act under certain limits. Should moral responsibility obey uniquely to regulative control, then a person could expect life to be either a succession of fortunate experiences, or some sort of Greek tragedy – even if determinism does not necessarily imply that we have 'destinies', meaning that our choices are inconsequential.

Continuing with the case of the mysterious scientist, his presence does not make any action unavoidable in a world that is completely indeterminist, while in a determinist world the presence of the mysterious scientist is superfluous. In this way, and to put it in Fischer´s words, the *Frankfurt Style* cases show that moral responsibility does not require regulative control. For that reason, even when there were no such regulative control, there would still be guidance control that does not require alternative possibilities such as when a person turns to the right with his car, even if he could not, due to technical problems, turn to the left[10].

If determinism excludes regulative control, but does not exclude guidance control, that is because moral responsibility is based on the capacity of the agents to control their acts: both in the capacity to answer to the acts of other agents, as well as the conducts that imply mechanisms

causal determinism being compatible with moral responsibility, apart from if causal determinism eliminates the access to relevant alternatives.

[10] Consider also the classical example of Locke (1992): suppose that a man is moved to a room while he sleeps. When he wakes up he sees a person he wants to see and with whom he wants to speak. Suppose, also, that he was locked up without his noticing, in such way that he cannot get out. When he wakes up, he will be happy to find the desired company, with whom he will decide to stay. That is to say, he will prefer to stay in there instead of going out. Locke wonders: Is this stay voluntary? And he answers that nobody will doubt that it is voluntary, even though, considering he has been locked up, it is evident that he has no freedom to decide whether he stays or leaves.

of rational deliberation. In the same way, John is responsible for many of his choices even though he is causally determined by the mysterious scientist. Indeed, guidance control refers to the mechanisms inherent to the agents to carry out an act, since it consists in a type of counterfactual dependency of the actions over reasons or motives. And, according to Fischer, a person can find reasons or motives even in a determinist world.

Fischer's ideas are based on the capacity to perceive oneself and to do things *in one's own way* - after all, guidance control is some sort of valuation of one's expression. For example, consider the moment before facing death – of course, if the world is deterministic, the *way* in which we die is determined as well as our reactions towards it. If one agrees with Heidegger (1977) and considers death as the last possibility, then it is not difficult to speculate over certain existential compatibilism, like Fischer. Death is the last possibility and we are causally determined to face it, but there is not one only way to do so. How it is done relies on the authenticity and autonomy of the agent. In this way, one is not responsible for his own death, but is responsible for the way in which he reacts before it[11].

However, the change of perspective that ranges from alternative possibilities to discourses over authenticity does not cover all the cases. Consider the psychological process of someone who is an addict against his own will, e.g. someone who wants to quit smoking but cannot do so. A smoker struggles against his addiction because he is aware of the health problems that it brings with it. But, in some point in his struggle,

[11] In this order, here could be considered cases that range from the different ways in which ill people react to a terminal disease, to the Socratic reaction that has been expressed in the dialog *Phaedo* (leaving aside, of course, the issue of the suicide).

he stops trying to quit. He decides that he cannot keep on struggling and becomes an addict to his own will. So he starts to think that, even though his addiction is detrimental to his health, it is not worth to live without it and keeps on smoking. Is it possible to say that, after losing the desire to escape his addiction, he has now acquired the freedom and the responsibility to continue smoking? Therefore, there are times in which living life according to each one's intentions or ways does not guarantee that one is responsible for his actions.

3. Betty and Benji Cases

Consider the following case, extracted from Mele (1995). Betty was a six year old girl who was scared of the basement of her house, especially when the lights were out. She did not understand why she was scared, since she knew that nothing bad was going to happen to her. Then, she believed that her fear was childish and developed a strategy to overcome her fear: to go down to the basement periodically until she was not scared anymore. Betty was in control of herself. That allowed her to have a strong personality that helped her whenever she had to make choices in her life. Betty is now in charge of a position with a lot of responsibility, in which all decisions depend on her. According to Mele, Betty's course of action built her character in such a way that she became the person she now is, if we presuppose that there are causal chains that start with the intention of the agents. Betty was autonomous, since she went down to the basement intentionally as a consequence of her own decision. Therefore, and according to Mele, Betty is responsible for not being scared of basements nowadays, as well as for her strong personality. Of course, Betty was a child and, as every other child, she was influenced by her parents. But her parents' influence minimizes neither autonomy nor merit to her attitude. Similarly, the addict to tobacco also plans a strategy to overcome his addiction. The success of the strategy will depend on his persistence and self-control. The attitude of the addict is, at the beginning, anarchical in an Aristotelian way – i.e. his intentional conduct opposes a better judgment – and, for that reason, he does not have control over himself. He tries to change it, puts his effort in it, even though he fails by falling back into his addiction. As times goes on, he starts to resign himself, to lose faith in himself. His judgments start to change in such way that he ends up considering his addiction as something that has to be enjoyed in life. All in all, his judgment and course of actions coincide. And again, can one say that the addict is now in control over himself? When one is capable of making up reasons for which one considers oneself responsible for one's actions, one is also capable of making up reasons for which one does not consider oneself responsible. So, this statement seems to suggest that we choose whether to be morally responsible or not according to the discourse or story we build of ourselves.

It can be considered that, in Betty's case, there is a begging the question because she already had a strong personality at the age of six. Waller (2011) compares this case with that of her twin brother Benji. Unlike Betty, Benji did not carry out any strategy to overcome his fear. Benji was less sure of himself and, either consciously or unconsciously, avoided going to the dark basement. Nowadays Benji has a weak character and a weak personality. He usually avoids responsibilities, since he had much less resources to face them than his sister. Perhaps Mele is right and such choices have affected them in their subsequent choices in life; perhaps a Freudian psychoanalyst sees in it the reason of many of their current attitudes. However, and comparing both stories, the problem does not lie in that each one of them had the personality that made them be who they are, but in why Betty did overcome her fear and Benji did not. Stating why leads to further causes, where Betty's capacity to face circumstances similar to Benji ends up being a matter of luck. In this way, Waller concludes that Betty and Benji already had several differences before assuming different positions to the same problem. She is not responsible for her strong character – like Benji is not responsible for his weak character – without her choices miraculously transcending their own causal histories. The differences in their developed characters can be recognized without appealing to any miraculous transcendence, assuming that they were the product of earlier differences, with respect to innate capacities as well as influences that are out of control and for which nobody is morally responsible.

The investigation of the past as an explanation of the present may result valuable to modify conducts or to understand why we do what we do. But, we have to take into account that we can always find a reason to be how we are. This is due to the large number of cooperating causes that go unnoticed and that are more important than what they seem to be. Furthermore, those who exclude cooperating causes, by means of explanations, so as to focus on a principal cause, do so according to some interest. In this way, someone arrives to a last word by excluding many possible last words (Laera, 2011). The main reason why Betty and Benji have different characters is referred to as being the product of certain narrative constructions that try to explain – according to certain explicit or implicit interests – why someone acts the way he does. This kind of reductionism is unavoidable.

The story told is always more a simple listing in a serial or sequential order of events, because the narrative organizes them into an intelligible whole that can excuse or blame someone for their actions and attribute moral responsibility. For example, there are stories that defeat presumptions of responsibility. These narrative constructions or stories can be called excuses and can apply in some cases but not in all. To be plausible, excuses must be found as socially acceptable. A murderer cannot evade his responsibility by telling a story about his genes. In these contexts the biological implication, as well as metaphysical implication, are irrelevant. Thus, agents can claim responsibility –or a lack of responsibility– for their actions depending on their relevant history, and this constitutes what we grasp simply as being responsibility.

It could be objected that all social narratives entail certain metaphysics. Suppose that the narrative mentions counterfactual situations: “When Betty was seventeen years old, she could have gone to Brandeis, but she chose to go to Harvard”. How are we going to interpret such counterfactual claims? We have to bring in some metaphysical idea of possibilities, to make sense of the narrative. However, the use of the subjunctive form does not imply the reference to metaphysical possibilities, and much less a hierarchy between moral narrative and metaphysical accounts. It can still conceive metaphysical possibilities as a mode of narration and establish a hierarchy with a moral narrative depending on the attribution context.

The narratives that imply moral responsibility can be built by focusing either in the third person or in oneself. It is possible to support Mele’s inner indeterminism by creating narratives focused in the intentional capacity of the agents in connection with their autonomous being. But it is also possible to take into account other cooperating causes to conceive them as principal causes, so as to focus the narrative in the environmental conditions and minimize the importance of the characteristic of being autonomous. Characterizing moral responsibility as a way of narrative explains why the reasons of an action are so versatile. Being versatile means that, hermeneutically, there is an intentional orientation when one is looking for responsibilities: a request is not a request, nor is a demand a demand; one can opt to say no.

Assume that somebody builds a narrative that includes the assumption that social order determines the conduct of the agents. According to this conception, if S was to commit a crime, the reason will not lie in the individual who executed the action, but ultimately in circumstances that do not depend on the agent –this could even serve as an extenuating circumstance. Now, another person changes the hermeneutic context and takes into account more proximate causes, such as the hate that the murderer felt for his victim. Or course, both are in disagreement, since they have different criteria of responsibility. One is a narrative going back to the criterion that the origin of every action is outside the agent – where the ultimate cause of committing a crime can be social injustice, inequality of possibilities or the personality of the agent, etc. The other resorts to a criterion that takes as an origin the autonomy of the agent to decide for himself. For such disagreements to be epistemically authentic, they have to share the same conceptual frame. That is to say, they have to share the subject they are referring to. The point in common is that both are inscribed within the frame of a narrative that includes, either explicitly or implicitly, judgments of moral responsibility.

When there are disagreements, the context of epistemic evaluation plays a decisive role for one narrative to prevail over the other. The evidence that supports propositions in which *S* is morally responsible (e.g. “John knows that *S* is responsible for...”) answers to recognition of the reason of the action. To the extent that, the final causes, the ones which exempt the agent from any moral responsibility, do not have a major influence. For instance, in criminal law: even though the agent is, to a large degree, determined by social order, he is also morally responsible for his actions. Even if the murderer were morally incorrigible, some moral evaluation would be attributed to him. However, these conditions have a binding influence if anyone attempts to explain the cause of the actions through psychology or sociology. And this is possible due to the fact that we have the large capacity to interpret the phenomenon of moral responsibility as a unit that entails reductions in contributory causes[12].

Nevertheless, not only does the reduction in causes require a conceptual frame in keeping with the past circumstances, but it also requires a narrative process oriented towards possible future circumstances and towards the power of prediction. One knows that the murderer who does not repent from his crime will probably kill again, because stories of him murdering someone may be created and they can be conferred a certain degree of truthfulness.

When the degree of truthfulness is too high, i.e., that there is a great expectation for him to kill, then the story becomes a prediction about the future. But, to what extent can we predict the result of our actions? Betty foresaw that she would overcome her fear of darkness in the basement with the strategy of going down periodically. Yet Betty could not foresee what kind of person she would be when she made that decision, just as Benji could not foresee the long term consequences that would arise from ignoring the problem. Therefore, even when Betty and Benji shaped themselves when they were six years old, they did not have the intention

[12] Cfr., Willaschek (2010)

of being who they are now. Consequently, they seem not to be responsible for that. Bearing this in mind, the search for responsibilities is measured with a double standard: when the story resorts to history, a precise fact which is distant in time is usually found as causally important in order to explain why things happen; on the other hand, when there is an attempt to find out the future consequences of the actions, the practical reason is often limited to paying more attention to the short term effects than to the long term effects.

As seen in Betty's case, a story is built about responsibility in which the challenge of the basement *was* the key to shaping her personality, but a story is not built about responsibility in which the challenge of the basement *will be* the key to shaping her personality since, in the latter case, there is nothing similar to a deliberate intention. In fact, if it is argued that in a deterministic universe we ought to blame people who blame others, we thereby assume compatibilism since it could not be fair to blame the blamers otherwise.

4. Story and Future Consequences

Strawson (1994) suggests that there is a requirement of ultimate responsibility that cannot be met and that is an essential condition in order to establish that actions are morally responsible. According to Strawson, actions entail true responsibility when they are performed by virtue of a reason of the agent that causes them. If causal chains ‑ which range from our desires, beliefs and values, up to our interactions ‑ were built at random, without a basis of rules or epistemology whatsoever, nor by virtue of some kind of control that is external to the agent, then there would be no place in which to search for any kind of responsibility. Yet, if actions and reasons depended on the agent's own abilities, they should be chosen by principles for which, in turn, he should be responsible by other means of choice, and so on, *ad infinitum*.

The deterministic idea, as well as the idea of a complete indeterminism, rests on the fact that, if the series of causes should be followed until their ultimate source, it would be clear that our interactions are out of our control[13]. The moral determinists and indeterminists conclude that it is unfair to punish some and congratulate others only because of their behavior. Ultimately, the abilities for good/bad behavior are the result of either a transcendental future or of the goddess of fortune, which gives no grounds for moral justification. However, this conclusion presupposes certain compatibility since it would be pointless to talk about justice: both punishment and merit would also be determined by causality. Assuming one is not dealing with an extreme determinist nor with a complete indeterminist but with a skeptic, maybe like Strawson, then the construction of moral stories that attribute responsibilities may be arbitrary. Arbitrariness consists of the establishment of where in the causal chain one stops searching for responsibilities. If moral responsibility is to be thought in proximal terms and in a specific context, it is only because certain punishments are fair or unfair only when one disregards the ultimate source of all blame. One may blame someone for not going to work because he fell asleep, but one cannot blame someone for not going to work because he is sick. Falling asleep does not depend on one's free will, just as being sick; however, responsibilities are very different in one case and the other. Someone may argue that he should have gone to sleep earlier the night before and that he should have set

[13] See Keane (1996)

the alarm, but it is more intricate to build a story in which one should have avoided the disease. While both cases and elements are out of the agent's control, the story about responsibility will have the same shape: "had he done such thing and such other thing, then...", the difference is that, in the case in which one falls asleep it works, but in the case of the disease it is more difficult for it to work.

If neuroscience or the laws of nature identify the descriptions of responsible actions with a more basic type of description, then one could very well eliminate the story of moral responsibility in favor of another kind of story, whose vocabulary will be comprised by the physical properties of the brain. The problem with such eliminativism is whether it is possible to find said identification and, even if it is found, whether both stories serve the same function – assuming also that one is able to specify which function corresponds to moral responsibility.

Disregarding the eliminativist thesis, it is still difficult to estimate up to which point one should investigate in the causal chain, since one may investigate enough as to commit to the explanation of moral responsibility of our behaviors, just as one may investigate enough as to leave it aside. It is a matter of whether one should arrive to the sources that are out of one's control or not when causal explanations are sought. Imagine that Betty goes to the casino and she wins a lot of money, and then she decides to give half of that money to charity. Betty's decision may be said to be worthy of praise since she could have very well kept the money; and, at the same time, it may be said that it is a matter of luck, since she could have not won and she could have not had the possibility to give anything to charity. Therefore, there is a difference between, for instance, not being able to stop smoking and choosing to light up a cigarette for the first time. This difference persists even in a deterministic world. In the first case there is no control over the behavior; in the second case, one is assumed to choose. Regarding lighting up a cigarette for the first time, one can of course ask, to what extent is it really a choice? Lighting up a cigarette for the first time may be the consequence of peer pressure among colleagues (especially if one is talking about a teenager) plus a weak personality, etc. The answer to this kind of question is based on the idea that one is responsible for one's actions to the extent that, in the story that was built, some responsibility is taken, whether for oneself or for others.

Think about Milgram's (1963) famous experiment in which responsibility may lie in the authority of the scientist as well as in the "master" applying the discharge. Whoever applies the discharge may build a story that may exempt him from responsibility, while an observer may arrive to the conclusion that his behavior is immoral. The "master" may claim that he was only obeying orders and that he trusted the authority of the scientist; he could also affirm that "they know what they are doing." However, the observer may claim that, in spite of the pressure of the scientist, the person applying the electric shock, as an autonomous being, should have behaved, ultimately, in a different way. Beyond the surprising results that arose from the experiment, the idea behind this point is to indicate that one may justify one's actions in many different ways and that responsibility is not a matter that is independent from the story.

No one knows if we are causally determined and, even if we are, the truth is that we behave and evaluate ourselves morally as if we were not. This is so even if the systematic approach of determinism, whether metaphysical, naturalistic or environmental, were believable[14]. For example, a drunken person behind the wheel has no control over his actions, but that does not mean that he is not responsible if he hits another person, even if the source of his alcoholism were child abuse. Regarding blame and punishment, the degree of control over our actions is supported by a close responsibility that is vital for evaluative attitudes. If responsibility depends on the rational construction that conceives the actions of the driver as intentional, it is due to the fact that said construction entails the desire of truth in a counter factual judgment: “he could have avoided drinking when he was supposed to drive.” In such cases, responsibility lies in rules that seek to guarantee people’s safety. Therefore, it does not entail the search for an ultimate level of control; in the end, completely indeterministic conclusions are considered in this search.

Moral responsibility is not only a product of the construction of the story of past actions, but also the ability to teleologically evaluate, as correct or incorrect, possible future actions. To a certain extent, one can know the future, inasmuch as possible worlds may be represented and, from that, actions may be morally judged. Likewise, moral responsibility is also settled in the motivation of actions when they serve as a starting point for stories that predict future consequences with a certain degree of probability. This hope of achieving practical results that have been predicted guides most of our decisions. However, the ultimate consequence of our actions is a complete uncertainty, just as it is the ultimate cause of our actions.

Nevertheless, human behavior may be retrospectively evaluated by virtue of results that were not predicted by the agent and past

[14] In this respect, I agree with Hoyos (2009), for whom the subject of human freedom is mainly relevant in the scope of social philosophy and not in metaphysics.

explanations and causes may be constructed through actions or omissions so as no reproaches may arise, whatever the result may be. In fact, evaluative expressions may be justified inasmuch as they are characterized as intentional, even if they are not. Following Frankfurt (1969, 1971), there are circumstances in which coercion does not limit the responsibility of the agents. Frankfurt maintains the general idea that someone is capable of being morally evaluated, whether negatively or positively, for his performance, even if it was neither intentional nor deliberate.

The actions not only include direct personal behavior, but also the results and the consequences of what was directly done. For instance, by pulling the trigger of a gun, one can predict a bullet shall be fired, and that it may kill someone. In that case, direct personal behavior consists of moving the hand and the finger with which one aims the gun and pulls the trigger. In fact, this may be the only event one tries to deliberately produce: one may not want to fire the bullet or that someone may be reached or fatally injured by the shot. However, the direct personal behavior of aiming the gun and pulling its trigger represents only a part of what, in fact, was done. There is another part that has to be taken into account and that refers to not analyzing the mediate or immediate future consequences of the action of holding a gun. When someone has a gun in their home he is responsible of omitting, as a future pertinent possibility, the fact that said gun may be accidentally fired.

The story of moral responsibilities for actions or omissions is associated with the control over future consequences since, when the success of predictions is pointed out, the unspoken agreements with other agents that made it possible is omitted. The circumstances that condition the state of future things are trimmed in order to make the intention of the subject of the action more relevant. Therefore, taking the above mentioned example into account, one is responsible of not performing the action of considering as a future pertinent possibility the fact that the gun may be accidentally fired, even though that judgment is made *a posteriori*, since the person making that judgment constructs a story in which the subject of the action should have controlled the consequences of his actions.

Given that there are expectations that arise from the deliberation about future consequences, one may narrate the counterfactual aspects of one's actions and attribute intentionality to them. Consequently, ***S*** is responsible for a future action when he has the possibility to choose alternate actions that lead to other actions. Or he may rethink said action based on actions in order to arrive to the state of things: ***S*** is responsible, through one or several present actions, for a future state of things, when the consequence of the choice of his action is considered to be a relevant alternative. For instance, people who unload toxic waste into the catchment area of a river, even when it can be avoided, are responsible for the contamination of said catchment area in the future, if this consequence is considered to be, in the story, as a relevant alternative to the action (Campbell, 1997).

One is not able of teleologically evaluating every variable of one's decisions, although it is presumed that one is able of evaluating the more relevant ones. However, if this was so, should one not also be responsible for what one considers to be relevant?[15] While ramifications of the effects of the action always exceed the foreseeable consequences, we are willing to apply intentional properties to unpredictable long term

[15] One may even think of Kant, and that the impossibility of a correct evaluation of every effect of an action, that is to say, the impossibility of a complete teleological evaluation, gave rise to the categorical imperative.

consequences. Therefore, the relevance of the possible consequences will also depend on the degree of interest of the agent in his counterfactual deliberations, without excluding that the particular values and the ability of the calculations may always lead to something unexpected (Cf. Mulligan, 2006)[16]. Consequently, due to the fact that there is no ultimate control over what is done since every action is not more than the development of a *given* thing, there is no moral responsibility in the profound sense of the term[17]. However, while the lack of an ultimate control of our actions and the degree of interest of the agent in predicting relevant future consequences cannot be defined, the story of moral responsibility assumes a compatibility that is partially inevitable since said stories are centered in the near control that *should have been* considered to be relevant. Even though a hardly profound type of responsibility is worth saving, the moral experience works because it is not based on theoretical reflections about its nature.

5. Overall Conclusions

Personal responsibility regarding one's own future, in order to achieve certain purposes, entails the responsibility regarding the future of others. Yet, even if one cannot control how others influence one's decisions, this does not mean that one is not the owner of or responsible for the success or failure of one's own decisions. Our stories about moral responsibility are fundamentally compatibilist since judgments on the attribution of

[16] In fact, Nagel (1991) defended the notion that agents may be morally responsible for those actions they inadvertently produce or those actions that do not have an explicit intention.

[17] Such as Smilansky (2003) has maintained.

responsibilities are based on the ability of producing counterfactual statements. In this sense, throughout this paper certain ambiguity among the distinction between the rules and moral responsibility can be noticed. However, this leap between one aspect and the other occurs since the evaluations on the description of *what is* is determined by the context, just as the evaluation or the acceptance of *what must be*. The reason for this is that moral reasons, whether they constitute a description or an evaluation criterion, are context-sensitive. In certain contexts, narratives concerning the ultimate causes of a piece of behavior are constructed. In other contexts, other narratives or stories with a more limited range of causes are built and interpreted. Thus, certain moral responsibility judgments are incompatibilist while others are compatibilist.

Different stories, some about *what it is* and about *what must or should be*, imply considerations, inquiries, and different depths when it comes to the evaluation of moral responsibility. Therefore, inquiries about the responsibility of the principal agent of the case reach deep speculations about his past, his genes, or whether he is determined to act in a different manner or not and, in other occasions, these depths are not reached; the responsibility is rather attributed only based on certain considerations about the will of the agent, his efforts and the recognition of the action as his own. Then, there are stories that place the subject in the place of someone who complies with certain normative standards in order to be responsible and there are stories that do not in spite of the action being the same[18].

[18] For instance, Sher (2009) has suggested that one is sensitive to the attributions of moral responsibility if the agent is conscious of the moral value of the action at the time in which he acts or not. On the other hand, others consider that it is his training or education and if those prevent his autonomy or not (e.g., Haji and Cuipers, 2008).

In other words, on the one hand, when it comes to looking for responsibilities, the causal chains stop at the near causes, both in relation to the reconstruction of the past and in relation to future consequences. The search for profound or long term causes is a matter of theoretical activity that has nothing to do with prizes and punishments in practice. On the other hand, if determinism were true, it would be possible to have enough knowledge so as to predict the shape of the future without failing. Yet, human beings lack the pertinent knowledge and the necessary intellectual abilities, which means that the fact that we are not able to predict the future constitutes no evidence of the falseness of determinism. It does evidence the possibility of compatibilism since, as the future is unknown, stories around the agent's non-executed possibilities can be established, even if such speech is also determined.

The story with judgments on moral responsibility depends on the context of attribution. Like so, in a philosophical or theoretical context where the ultimate sources of our actions are sought, Betty may consider herself as not being responsible for her success, but in the practical context of our relationship with the world, Betty is responsible for her success. Therefore, if this notion is followed, moral responsibility is compatible with determinism. So, is it fair that Betty is rewarded for her decision? Is it fair that Benji is punished for his decision? The answer to both questions depends on the causal story one builds. If the causal demand is high, that is to say, that one tries to reach the ultimate source of the action, then it is possible to reach an explanation in which neither of them is responsible for the way they are. But this demand seems to be more philosophical or theoretical than practical. In the context of everyday life, the attribution of responsibility is quite simple whereas in more specific contexts, the search for responsibilities becomes more complicated. As one approaches the ultimate sources of responsibility, it dissolves among skeptical reflections.

References

Berofsky, B. (1995). *Liberation from Self.* New York: Cambridge University Press.

Berofsky, B. (2006). "The Myth of Source", *Acta Analytica 21*, pp. 3-18.

Campbell, J. K. (1997). "A compatibilist theory of alternative possibilities", *Philosophical Studies 88*, pp. 319-330.

Chisholm, R. (1966). "Freedom and Action" in K. Lehrer (ed.), *Freedom and Determinism*, New York: Random House, pp. 28-44.

Fischer, J. M. (2006). *My Way: Essays on Moral Responsibility*. Oxford: Oxford University Press.

Frankfurt, H. (1969). "Alternate Possibilities and Moral Responsibility", *The Journal of Philosophy 66*, pp. 829-839.

Frankfurt, H. (1971). "Freedom of the Will and the Concept of a Person", *The Journal of Philosophy 68*, pp. 5-20.

Haji, I., & Cuypers, S. (2008). *Moral Responsibility, Authenticity and Education*. New York: Routledge.

Hawthorne. (2001). "Freedom in context", *Philosophical Studies 104* (1), pp. 63-79.

Heidegger, M. (1977). *Sein und Zeit*, Frankfurt am Main: Klostermann.

Hoyos, E. (2009). "El Sentido de la Libertad", *Ideas y Valores 59*, pp. 85-107.

Inwagen, P. v. (1983). *An Essay on Free Will*. Oxford: Oxford University Press.

Keane, R. (1996). *The Significance of Free Will*. New York: Oxford University Press.

Laera, R. (2011). *Los desvíos de la razón: el lugar de la facticidad en la cadena de justificaciones*. Buenos Aires: Miño y Dávila.

Lewis, D. (1983). "New Work for a Theory of Universals", *Australasian*

Journal of Philosophy 61, pp. 343-377.

Locke, J. (1992). *Ensayo sobre el entendimiento humano*. México, D.F : F.C.E.

Mele, A. (1995). *Autonomous Agents: From Self-Control to Autonomy*. Oxford: Oxford University Press.

Milgram, S. (1963). “Behavioral Study of Obedience”, *Journal of Abnormal and Social Psychology 67*, pp. 371-378.

Mulligan, T. (2006). *Future People*. Oxford: Clarendon Press.

Nagel, T. (1991). *Mortal Questions*. Cambridge: Cambridge University Press.

Pereboom, D. (2001). *Living Without Free Will*. Cambridge: Cambridge University Press.

Sher, G. (2009). *Who Knew? Responsibility without Awareness*. New York: Oxford University Press.

Smilansky, S. (2003). “Compatibilism: The Argument from Shallowness”, *Philosophical Studies 115*, pp. 257-282.

Strawson, G. (1994). “The Impossibility of Moral Responsibility”, *Philosophical Studies 75*, pp. 5-24.

Wallace, R. J. (1994). *Responsibility and the Moral Sentiments*. Cambridge: Harvard University Press.

Waller, B. (2011). *Against Moral Responsibility*. Cambridge: MIT Press.

Waller, B. (1990). *Freedom without Responsibility*, Philadelphia: Temple University.

Willaschek, M. (2010). “Non-Relativist Contextualism about Free Will”, *European Journal of Philosophy 18*(4), pp. 567-587.

University of Barcelona
rodrigolaera@gmail.com

철학적 분석 31 (2014) pp.151-181

Disattendability, Civil Inattention, and the Epistemology of Privacy

Axel Gelfert

The concept of privacy is intimately related to epistemological concepts such as information and knowledge, yet for the longest time had received only scant attention from epistemologists. This has begun to change in recent years, and different philosophical accounts have been proposed. On the *liberal model* of privacy, what privacy aims at is the protection of individuals from interference in personal matters. On the (more narrowly epistemological) *informational model*, privacy is a matter of limiting access to (or maintaining control over) certain types of information. Furthermore, it is sometimes claimed that privacy aims at preventing the formation of (potentially negative) judgments by others. This paper compares and contrasts the various approaches and identifies a number of shortcomings. It then outlines an alternative account, the *disattendability/civil inattention model of privacy*, according to which what constitutes a breach of privacy is neither the acquisition of new information *per se*, nor the formation of judgments by others, but the fact that undue attention is being paid to routinized (or otherwise unobtrusive) aspects of the target's everyday life. This account, it is argued, is explanatorily superior to its competitors, in that it accounts both for the cultural contingency of privacy conventions and for the emergence of new threats to privacy (e.g. from electronic surveillance). The paper concludes by reflecting on remaining tensions within the proposed account of privacy and by commenting on its social ramifications.

Keywords disattendability, epistemology of privacy, privacy, social epistemology

1. Introduction

It is something of an oddity that the concept of privacy – which hinges on epistemological concepts such as information, knowledge, and its communication – has, for the longest time, been discussed not by philosophers, but by legal scholars. Partly as a result of this, there seems to have been less emphasis on providing a rigorous analysis of the concept of privacy, and more interest in developing a cluster of somewhat imprecise, but 'serviceable' legal considerations. In recent years, philosophers – especially those with an interest in social epistemology – have begun to redress this imbalance (see, e.g., [Fallis 2013] and [Matheson 2007], and references therein), yet philosophy still has a long way to go if it is to catch up with the legal tradition of thinking about privacy. This paper is intended as a contribution to this philosophical project and is organized as follows. In Section 2 ('Models of Privacy'), after briefly rehearsing the origins of the privacy debate in legal theorizing, I will quickly turn to two contrasting classes of philosophical approaches, which I shall refer to as the *liberal* and *informational* models of privacy, respectively. Their validity will be tested against several scenarios (Section 2.3.), and we will see that their shortcomings motivate a third type of account (the *immunity model*), which posits that privacy seeks to protect us from the judgments of others (Section 3). In Section 4, I will introduce an alternative account, which takes as its starting point Erving Goffman's twin notions of *disattendability* and *civil inattention*. In a nutshell, while the principle of disattendability demands that, in public settings, I am not to be obtrusive, civil inattention requires of us to suspend specific attention to others and their behaviours in various circumstances. As I shall argue, a model of privacy based on these twin concepts is explanatorily superior to its competitors and holds out the promise of being able to account for both cultural diversity and the changing face of privacy in an age of electronic surveillance. I will conclude (Section 5) by commenting on some remaining issues, notably the worry that the proposed *disattendability/civil inattention model of privacy* might seem to unduly privilege social conservatism, and by reflecting on the prospects of privacy as context-dependent and domain-specific phenomenon.

2. Models of Privacy

As mentioned in the Introduction, the legal tradition of thinking about privacy predates its philosophical analysis. (See, for example, [Solove 2008] and [Wacks 2010], and references therein; for an early philosophical treatment of the topic, see [Thomson 1975], who argues that violations of privacy are reducible to violations of other, more fundamental rights.) More specifically, the legal tradition can be traced back to an influential essay by Samuel Warren and Louis Brandeis, published in the *Harvard Law Review* in 1890, under the title 'The Right of Privacy'. Warren and Brandeis's essay attempts to give substance to privacy as the legal 'right to be let alone' and was motivated by the intrusions of the yellow press into the lives of (often famous) individuals. More specifically, Warren and Brandeis reacted to the – then novel – technology of photography, arguing that 'the law must afford some remedy for the unauthorized circulation of portraits of private persons' (Warren and Brandeis 1890: 195).

In spite of its considerable influence in the legal domain, Warren and Brandeis's essay makes for a somewhat disappointing read, at least for the philosophically inclined reader, since its authors are not so much concerned with the *concept* of privacy, but instead aim to derive a generalized right to informational privacy from elements they take to be implicit in the common law tradition that had always afforded persons and property some degree of legal protection. Robert Post voices a similar sense of dissatisfaction:

> The prestige and enormous influence of the [Warren and Brandeis] piece creates expectations of sweeping vistas and irresistible arguments. But, setting aside the rhetorically powerful (and often quoted) passages of complaint against the irresponsibility of the press, the article offers instead a technical and rather dry exposition of the legal rights of unpublished authors and artists. (Post 1991: 647)

Although ostensibly aimed at overcoming a narrow focus on property right by establishing a broader right to privacy, Warren and Brandeis nonetheless remain indebted to the property-based conception of rights – for example, when they argue that the 'fiction of property' can be preserved by recognizing that 'it is still true that the end accomplished by the gossip-monger is attained by the use of that which is another's, the facts relating to his private life, which he has seen fit to keep private' (Warren and Brandeis 1890: 205). It should be clear that such conceptual contortions are largely the result of trying to shoehorn a complex and novel question – how to regulate the flow of information in an age of sensationalist reporting – into a legal tradition which, at the time, was lacking in descriptively and normatively adequate conceptual resources. For this reason, in the remainder of this section, I will not be looking towards the legal tradition for guidance, but instead will draw on two sets of considerations that emerge from political liberalism and the epistemology of information, respectively.

2.1. Liberal models of privacy

Broadly speaking, liberal models of privacy seek to identify a domain of individual behaviours, beliefs, and activities, which – in virtue of their private character – should be protected from regulation (or persecution) by the state and, more generally, from unwarranted intrusion by others. The classic expression of the general sentiment underlying liberal models of privacy can be found in John Stuart Mill's *On Liberty*, who already recognized – in a manner that must surely seem prescient to any 21st-century reader familiar with such phenomena as cyberbullying – that society can be more oppressive than state regulation:

> But reflecting persons perceived that when society is itself the tyrant – society collectively over the separate individuals who compose it – its means of tyrannising are not restricted to the acts which it may do by the hands of its political functionaries. Society can and does execute its own mandates: and if it issues wrong mandates instead of right, or any mandates at all in things with which it ought not to meddle, it practices a social tyranny more formidable than many kinds of political oppression, since, though not usually upheld by such extreme penalties, it leaves fewer means of escape, penetrating much more deeply into the details of life, and enslaving the soul itself. (Mill, *Complete Works* 18:219)

On the one hand, liberalism calls for real protection from substantial interference; on the other hand, it also calls for self-restraint in our interactions with others, even in matters that stretch the notion of what should be considered the private realm, such as conflicting fundamental beliefs that arise from competing conceptions of the good. As Joshua

Cohen notes, ‘cultural liberalism requires that, as a general matter, we steer clear of controversial topics about which we cannot expect to reach agreement and that do not demand a collective decision’ (Cohen 2009: 318). This, as Cohen recognizes, gives rise to the thorny question of how to delineate a genuinely private realm from that which is fit for public deliberation and scrutiny, given that it is hard to see ‘how we could possibly identify the private arena with the family, or with the economy, or with any arena of social life, identified – either spatially or institutionally – prior to normative political argument’ (Cohen 2009: 312).

Since the present paper focuses on the epistemology of privacy, I shall not delve any deeper into its normative and political dimensions, important though these are. It is worth noting, though, that an epistemological analysis of the concept of privacy is bound to have repercussions for the substantive normative debate, insofar as we first need to develop a theoretical understanding of when the possession of certain types of information constitutes a breach of privacy in the first place.

2.2. Informational models of privacy

Unlike the liberal model of privacy, informational models are not primarily concerned with substantial interference with individual liberties, but instead link breaches of privacy more directly to the possession of private *information* (which may or may not have been illegitimately obtained). A good illustration of the informational model is this definition of privacy, given by W. A. Parent:

> Privacy is the condition of not having undocumented personal knowledge about one possessed by others. A person’s privacy is diminished exactly to

> the degree that others possess this kind of knowledge about him. (Parent 1983: 269)

Parent immediately points out that this definition pertains to 'the condition of privacy, not the right to privacy' (ibid.) and, indeed, it is the assumption of epistemological primacy of the former – that is, of the conditions that need to be in place for privacy to obtain – which distinguishes the informational model from the liberal model (whose epistemological underpinnings are often left implicit, even while its proponents aim to establish a right to privacy). For a breach of privacy it is not required that personal knowledge about a person be used against him, but rather that such information is obtained by an unauthorized party without his consent. Klemens Kappel makes a similar observation when he writes: 'According to the generic epistemic theory of informational privacy, privacy depends inversely on epistemic access.' (Kappel 2013: 185).

However, the generic theory cannot be the full story. For one, there is some ambiguity in how we should understand the term 'access'. If access is equated with mere availability – that is, if it is understood in dispositional terms – the generic theory would likely overgenerate cases of privacy being breached. For example, when a file containing a patient's medical history is left unattended, so that in principle anyone could access it, we may not wish to speak of a privacy breach unless someone *actually* accesses the patient's information. (Having said that, there may be good practical reasons for defining 'privacy' in such an inclusive manner, not least because it might encourage good data protection habits.) Furthermore, it does not sound quite right to say that one's condition of privacy is compromised exactly in proportion to the number of individuals who either have access to it, i.e. *inversely to epistemic access* (as the generic theory asserts), or ‘exactly to the degree that others possess [the information in question]’ (as Parent puts it). Even on the informational model, the term ‘privacy’ should amount to more than just a convenient label for the degree to which other individuals have access to, or possess, certain types of information – that is, privacy should be more than a mere ‘numbers game’.

This last point is echoed by Masahiko Mizutani who notes that a purely dichotomous division between public and private would overlook the fact that privacy need not be compromised if information is willingly shared with select confidants. For example, ‘with regard to certain types of information about myself’, I may ‘convey everything to my wife, only the summary to my friends, and nothing to strangers’; yet, ‘in reality there may be things I am willing to tell my friends but not my wife, or secrets I am willing to reveal to passersby during my travels precisely because they are strangers who do not know me’ (Mizutani 2012: 610). It seems to me that, in the event that I share a (perhaps embarrassing) personal story with, say, a fellow train passenger – whom I will never meet again – not only is it the case that, a fortiori, my *right to* privacy is not violated (given that it was my free decision to tell the story), but my *condition of* privacy is not compromised either, even though the number of individuals with access to the embarrassing information has increased. Any account of privacy should be able to explain why access *sometimes* compromises privacy, and sometimes doesn’t. If in-principle access sufficed to undermine privacy, then it would seem that we would have to restrict ‘true privacy’ to those private mental items to which only the individual cognizer has access. Lest we opt for such a restrictive usage, which would be needlessly revisionist of social, linguistic and legal practice, we should reflect on which *kinds* of access undermine privacy and which do not.

2.3. Some test cases

Let us consider a few cases that will illustrate the limits of the (access-based) informational model. Jeffery Johnson (1989) offers two instructive scenarios, which will form the basis for the discussion in this section.

> Bugged phone. 'You place a bug on my phone in order to overhear the most intimate details of my conversation with my mistress. We are too smart for you; we have already agreed to never discuss intimate things on the telephone [...]. Your tap reveals nothing; you have gained no new personal information. You certainly have, however, violated our privacy. Such cases cannot plausibly be dismissed as mere *attempted* violations of privacy.' (Johnson 1989: 161)

As Bugged phone makes vivid, *actual* acquisition of personal knowledge is not required for a breach of privacy to have occurred since, as the case has been described, no personal information is passed on. Perhaps whoever intercepted the phone line was hoping for more, but intentions on the part of third parties have no place in the informational model. To the extent that an instance of unauthorized access has occurred, this pertains to the *mode of communication* – wiretapping into someone else's phone line – not to any sensitive information itself. Perhaps, then, what undermines the condition of privacy in the case at hand is the fact that, had it not been for the unusual precautions taken by the two lovers, personal information *might have easily* been acquired. This, too, however, is not easily accommodated by the informational model, at least not in its 'generic' formulation (see previous subsection).

But even when no *new* personal information can be acquired – for example, because the intruder already has all the information she could possibly hope to obtain – this does not preclude the possibility of a breach of privacy. Consider the following case:

> Voyeuristic Sarah. 'Sarah peeks through my bathroom window in order to admire my naked body. [⋯] Suppose that Sarah is both my physician and a frequent guest in my home. As a guest she knows all about the vivid rose paint job in my bathroom; and as my physician she knows only too well about the sorry state of my waistline. She nevertheless clearly violates my privacy when she peers in my window.' (Johnson 1989: 161)

In the case as described, Sarah does not – and, indeed, could not hope to gain – any new information: as a frequent house guest, she knows all about the bathroom, and as her target's physician she has probably seen more of him than she could reasonably expect to glimpse through the bathroom window. No new information is being accessed, not even potentially, yet the *manner in which Sarah accesses (known) information* clearly constitutes a breach of privacy.

One attempt to amend the informational model might be by positing control over one's personal information as a desideratum of privacy. On what one might call the *control model of privacy*, privacy would be determined by the extent to which we can control what others come to know about us. As an example of a control-based account, consider this statement by Charles Fried:

> Privacy is not simply an absence of information about us in the minds of others; rather it is the control we have over information about ourselves. To

> refer for instance to the privacy of a lonely man on a desert island would be to engage in irony. The person who enjoys privacy is able to grant or deny access to others. (Fried 1984: 209–10)

The ability 'to grant or deny access to others', however, depends on many contingent factors, and it is far from clear that we should only speak of 'privacy' when such control can be reasonably assumed. Consider this example discussed by Martijn Blaauw:

> Diary. 'John is a very private person. All the truths about himself that he finds important are not known to any others. Yet today, before leaving for work, he left his diary open on his desk. His cleaning lady could open the diary and read its contents, thus acquiring knowledge of John's personal propositions. And being in meetings all day, John has no means to go home and prevent this.' (Blaauw 2013: 175)

As Diary makes clear, John, because he has failed to place his diary under lock and key, could not possibly prevent his cleaning lady from reading what he has written and, in this sense, is *unable* to grant or deny access to it. His (temporary) inability to control who has physical access to his diary, however, does not in and of itself undermine the condition of privacy – at least 'so long as the cleaning lady hasn't perused his diary' (Blaauw 2013: 175). It would seem, then, that the control model of privacy ends up collapsing into the generic informational model – according to which privacy is a matter not of access or control, but of the actual possession of information – in which case, however, it would also inherit the latter's problems.

Alternatively, we may wish to shift the focus from the ability to restrict access to information to the basis on which granting or denying access, though not always successfully, may be deemed legitimate. In other words, perhaps we should inquire into what it is that (attempted) restrictions on the flow of private data and personal knowledge aim to achieve. One idea, to be developed in the next section, will be that restrictions on the spread of personal knowledge serve the purpose of shielding an individual from the judgment of others. Before turning to this account, however, it will be useful to reflect briefly on what we mean by 'personal' information – given that, in common parlance, this label suggests that we are dealing with sensitive issues which may, for example, be important to a person's self-image or sense of identity. As Kappel notes, we usually 'think of sensitive facts as including facts about our religious denomination, health, sexual preferences and certain parts of our lifestyle'. In addition, 'sometimes one's political views are regarded as sensitive, as are facts about one's financial affairs', or even 'one's true views about colleagues and friends' (Kappel 2013: 183).

Even within societies that are culturally similar (say, Western Europe and the United States), there is often considerable variation in what is conventionally regarded as private information – revealing one's salary may be taboo, whereas political affiliations may be freely volunteered by most people, and vice versa in another country – and this diversity increases as one crosses cultural boundaries. (On this point, see [Mizutani, Dorsey, and Moor 2004].) Thus, Parent argues that a full explication of the notion of privacy

> requires that we clarify the concept of personal information. My suggestion is that it be understood to consist of *facts* about a person which most individuals in a given society at a given time do not want widely known about themselves. (Parent 1983: 269-270)

While the notion of ‘personal’ (or ‘sensitive’) information retains much of its significance, recent technological trends – such as the rise of ‘Big Data’, involving the ‘mining’ of (individually insignificant) user data for clues as to, say, an internet user’s personal (e.g. sexual) preferences – may force us to rethink what kinds of data are relevant to the issue of privacy. What matters, of course, is not the mere *existence* of such technology, but its actual *deployment* as a way of gathering, collating, and aggregating otherwise highly distributed data points concerning individuals and their preferences. Much of the data that is being collected online, by social networks and advertisers, is individually mundane and boring, and not something that we would usually regard as constitutive of our personality, yet it may provide sufficient clues for an algorithm to make powerful inferences about some very personal aspects of individuals. As Mizutani puts it, ‘many of the problems relating to privacy in our current age are problems of a more latent nature; that is, regardless of whether privacy is actually or violated or not, what matters is the way in which our behavior is influenced by the mere feeling someone might be watching’ (Mizutani 2012: 610). As I will argue in Section 4, this idea can be made more rigorous by attending to the relational character of one person *attending to* another’s actions, preferences, and characteristics.

3. Privacy and the Judgment of Others

Towards the end of the preceding section I argued that we should inquire into what it is that attempts to restrict the flow (or accessibility) of private data and personal knowledge aim to achieve. As we saw in the various cases discussed above, privacy can be breached even when no personal knowledge is made accessible (Bugged phone), or when no *new* personal knowledge is acquired (Voyeuristic Sarah), yet no such breach needs to happen even when extensive personal knowledge is in plain view (Diary). It may be preferable, therefore, to focus not on the flow of information itself but on the uses to which it is put.

Johnson has suggested that one reason why we are especially concerned about personal information is that it may lead to unfavourable opinions that others form about us, due to their disapproval of certain types of behaviour (or certain facts about ourselves) that we normally try to hide from public view. As Johnson argues,

> all examples of privacy have a single common feature. They are aspects of a person's life which are culturally recognized as being immune from the judgment of others. (Johnson 1989: 157)

The basic idea is that, wherever there are cultural norms of privacy in place, societies have implicitly agreed to treat the requisite information as not fit for the purpose of assessing others on its basis. (Such collective normative agreement, of course, does not imply that no one will judge another person on the basis of personal information about him or her; indeed, it is precisely *because* many people are judgmental that personal information stands in need of protection!) Johnson's *immunity model of privacy* – according to which privacy is a matter of being immune from the judgment of others – echoes Mill's point about the oppression that comes with the experience of social disapproval. Where Mill argued that society 'practices a social tyranny more formidable than many kinds of political oppression', Johnson pinpoints the social mechanism by which society achieves this pervasive influence: through the *judgments of others*.

Johnson claims that his account fares better in explaining cases not covered by the liberal and informational models of privacy. Consider the following case:

> Golf on Sunday. 'I routinely play golf on Sunday mornings. I do not take precautions to hide this information. My neighbor in seeing me leave for the course every weekend does not violate my privacy. If, however, she takes it into her head to lecture me on my heathen ways and how my time on Sundays could be better spent in more spiritual pursuits, she has now intruded into a sphere of my life that is private. Her possession of personal information is not the issue; her negative judgment very much is.' (Johnson 1989: 161)

On the informational model, there is little that would mark out this case as a breach of privacy: no attempt is made by me to hide the information that I routinely play golf on Sundays, and my neighbour's gaining that information by looking out the window does not by itself breach privacy. Yet, it seems to me that, in the scenario as described, it is the neighbour's *lecturing me*, not her negative judgment *per se*, which infringes upon my privacy. In this sense, the Golf on Sunday case appears to be covered by the liberal model, and the specific contribution of judgments on the neighbour's part (except as the psychological cause of her unwarranted interference) remains unclear.

The immunity model (or, as Johnson prefers to call it, *judgment-of-others model*) also becomes implausible when applied to certain clear cases of privacy violations. Consider the case of Bugged phone, where the intercepted interlocutors had devised an elaborate system of rule to make it impossible to extract information. As Johnson recognizes,

> [t]here is a sense in which this counter-example to the information model could be marshalled against the judgment-of-others model. If the tap reveals no new information about our relationship, it is not immediately apparent how the eavesdroppers have formed any new judgment about us. They have certainly not, however, remained emotionally neutral. The very acts of monitoring my phone conversations implies a great deal of (emotional) interest in this aspect of my life. (Johnson 1989: 166)

This defence of the role of judgments on the part of the eavesdropper strikes me as deeply implausible. For the wiretapping of the phone to constitute a breach of privacy, no emotional investment on the part of the interceptor needs to be posited: Imagine the phone conversations were being recorded by a jaded surveillance specialist who 'has seen and heard it all' and who is deeply bored by (or, lest boredom be counted as an 'emotional investment', totally indifferent to) the affair, no matter how 'juicy' (or upsetting to others) its details may be. Surely, this would still count as a severe breach of privacy. Indeed, it seems to me that violations of the condition of privacy do not require the involvement of *any* human judgment. For example, the gathering of information could be outsourced to physical instruments (such as recording devices), and whatever analysis is performed may be carried out entirely by algorithms – which, as the whistleblower Edward Snowden has revealed, is the standard mode of operation of the United States' security agencies and their allies. Its automated nature, however, does not render such interception any less of a breach of privacy.

Two further worries about the immunity model merit attention, both of which concern the extent to which we can reasonably demand not to be judged by others. When discussing the Golf on Sunday scenario, Johnson argued that, whereas my neighbour's possession of personal information about my golfing habits ‘is not the issue[,] her negative judgment very much is’ (Johnson 1989: 161). To the extent that negative judgments may often form the basis of unwarranted interference by others – as in the case of the fundamentalist neighbour – liberal models of privacy may be able to accommodate such cases. Perhaps in order to distance his own position from liberal models, Johnson elsewhere argues that immunity is not limited to negative judgments only:

> One should not infer, however, that the concept of privacy only addresses the negative judgments of others. The professional model, who has no reason to fear a negative evaluation of her body, can still have her privacy violated by the voyeur. [...] When we claim immunity from the judgment of others, this applies to all judgments, positive, negative or neutral. (Johnson 1989: 163)

Yet, this notion of immunity seems too expansive by far: Surely it can neither apply to all areas of life, given that robust judgments concerning others are required in many contexts – not least for practical projects that depend on social cooperation for their success – nor can the mere subjective expectation of immunity from judgment be sufficient to ground privacy. In connection with the hypothetical case of a married couple’s public embrace being plastered all over the front page of a newspaper, Johnson writes:

> It is also, at least in this case, an area of your life in which you have a right to expect immunity from the judgments of others, even if the data had been correct. (Johnson 1989: 162)

Most of us, no doubt, would concur that the newspaper’s publication would constitute not only a violation of the right to privacy, but would also undermine the legitimate expectation that a state of privacy about a couple's personal matters be respected. But the case only serves to raise another question: Which are the areas of our lives where we can *legitimately* appeal to immunity from the judgments of others? One response would be to say that any answer to this question must be based on normative agreement about which areas of individuals' lives society deems worthy of protection – in which case we find ourselves back at Cohen's point that, 'prior to normative political argument', there is no telling which areas of life should enjoy privacy and which should be available for scrutiny. Beyond political and legal ways of managing privacy, however, the underlying question – 'What is it about certain ways of *judging others* that makes them (as opposed to other types of judgment) intrusions into other people's privacy?' – still awaits elucidation.

A second, related worry concerns Johnson's conclusion that all judgments that others form about us 'without our consent', as it were, are illegitimate. For this contradicts the core enlightenment idea of testing one's judgments against those of others – and, by extension, having one's judgments *contested* by others. (Immanuel Kant provides us with a clear expression of this idea when he remarks in his discussion of testimony as a source of knowledge that 'Man always wishes to test his judgment on others' and that it would be a form of 'logical egoism' to dismiss the need 'to compare one's opinions with those of other people'; see the discussion in Gelfert 2006.) Importantly, such a willingness to at least *consider* criticism should also extend to (some) *unsolicited* judgments: we would deprive ourselves of much useful criticism if we were to dismiss all unsolicited judgments of others as an illegitimate assault on privacy.

4. The Principle of Disattendability

Delineating the areas of life in which one can legitimately appeal to privacy in order to ward off unwarranted intrusion, unwanted attention, and unsolicited judgments of one's behaviour by others should be a central part of any philosophical account of privacy and cannot be relegated to an afterthought. Importantly, such an account should not only list the areas of life that various societies have deemed worthy of protection, but should also aim for some degree of unification, for example by identifying a common feature or mechanisms that tells us what would be objectionable about privacy intrusions in the various cases. (The immunity model proposes one such feature – the formation of unsolicited judgments on the part of others – but fails for the reasons discussed in the previous section.)

In order to develop an alternative model of privacy, I shall help myself to Erving Goffman's notion of *disattendability*. Disattendability, on Goffman's understanding, correlates with a form of *civil inattention* that we accord others in public places, across a range of acceptable behaviours. Unpacking this suggestion, as we shall see, will lead to important questions regarding the notion of acceptability as well as the nature of publicity. For the moment, however, let us follow Raymond Geuss in characterizing a *public* place as

> a place in which I can expect to be observed by 'anyone who happens to be there', i.e., by people whom I do not know personally and who have not necessarily given their explicit consent to entering into close interaction with me. (Geuss 2001: 58)

Precisely because, in public settings, we cannot rule out the possibility of our being exposed to others, we need to regulate the ways in which attention is being deployed in our dealings with each other. Civil inattention towards others is a way of recognizing that they have an equal claim to making use of public space and that our being co-present in the same space neither requires justification nor incurs special responsibilities towards each other. According to Goffman, it is through routine markers of civil inattention – e.g., temporary (brief) eye contact between strangers – that 'individuals exert respectful care in regard to the setting and treat others present with civil inattention' (Goffman 1971: 331), so that public order can be maintained.

At the same time as we accord civil inattention to others, public space is governed by a norm of disattendability: Since, in many public settings, we find ourselves surrounded by others who did not specifically choose to associate with us, their co-presence imposes on us the demand 'not to make undue claims on people's fears or trigger their embarrassment and disgust' (Miller 1997: 1999). Elevating this demand to the level of a behaviour-guiding maxim, we may then speak of the *principle of disattendability*, which is aptly described by Geuss as follows:

> The principle of disattendability states that in such contexts and places I am to be unobtrusive. That is, I am to allow the other whom I may encounter to disattend to me, to get on with whatever business they have without needing to take account of me. I am not to force myself on their attention. (Geuss 2001: 58)

The principle of disattendability can be made even more vivid by reflecting on some of its more blatant violations. Thus, Geuss recounts the story of Diogenes of Sinope, who

> was in the habit of masturbating in the middle of the Athenian market place. He was not pathologically unaware of his surroundings, psychotic, or simple-minded. Nor was he living in a society that stood fairly low on what we take to be the scale of our cultural evolution [...]. Rather, we know that the Athenians objected to his mode of life in general and to this form of behavior in particular. We know this because the doxographic tradition specifically records Diogenes' response to a criticism of his masturbating in public. He is said to have replied that he wished only that it were as easy to satisfy hunger by just rubbing one's belly. (Geuss 2001: 57)

Evidently, Diogenes' masturbating in public – being difficult to ignore when confronted with it – violated the principle of disattendability. Clearly, Diogenes should have limited this activity to a more private space. Perhaps, then, what motivates a distinction between the public and private realm – and, in turn, creates a need for shielding the latter from intrusion – is the fact that certain morally permissible actions and activities (and information pertaining to them) would violate the principle of disattendability *if* they were to be dragged into the public realm.

Can the principle of disattendability form the sole basis for an account of privacy? Recall that, as Goffman puts it, the principle serves to transform 'normative expectations into righteously presented demands' (Goffman 1963: 2) on others. This should raise some alarm bells for anyone who cares about protecting the private realm from undue moralization by others. After all, the principle of disattendability presents itself, in the first instance, as an imposition of a duty of restraint on the individual, who ought to behave in such a way as not to attract undue attention to himself. Yet, what is regarded as ordinary, routine, and normal in a given setting is heavily dependent on contingent cultural factors and conventions. It might seem, then, that privacy needs to be 'earned' by behaving in a conformist fashion. And indeed, critics have pointed out that disattendability is a socially problematic notion. As Margaret Urban Walker puts it:

> Civil disattendability shades off rather too readily into norms of 'respectability' that load hierarchical social arrangements, sheer prejudice, and socially sanctioned contempt for and exclusion of certain groups or their 'ways' from specific social locales. (Walker 2006: 124)

This is a legitimate worry, which I will address in the final section of this paper. It would be hasty, however, to dismiss the relevance of disattendability (and its correlate, civil inattention) to the project of developing a robust account of the notion of privacy. For, as we have previously seen, other proposed models of privacy also have their share of problems, and it will be instructive to see how the *disattendability/civil inattention model* fares by comparison. For one, it is important to keep in mind that disattendability is only one element of the model, civil inattention being its correlate: Where the disattendability principle demands that we behave ourselves so as to be civilly disattendable, civil inattention requires that we suspend specific attention to others and their behaviours or characteristics. It is this second demand – that we should accord others, who are simply going about their business and living their lives, the requisite degree of civil inattention, so as to not disrupt their routines, or the routines of public order in general – which may provide at least a tentative starting point for an alternative model of privacy.

In order to better assess the prospects of an account of privacy based on disattendability and civil inattention, let us turn to some of the test cases discussed in Section 2.3. Consider Voyeuristic Sarah, the case of the voyeuristic family friend and doctor. As we saw earlier, Sarah gains no information by watching her target getting undressed in his bathroom, since she is already well-acquainted with all the relevant facts. And, as Johnson notes, even if new information were to be gained 'my outrage at this violation of my personal privacy is not focused on the knowledge Sarah gained' (Johnson 1989: 161). Johnson concludes – rightly, in my view – that the informational model fails to adequately account for this breach of privacy. How does Johnson's immunity model fare? In line with its guiding idea, according to which privacy is a matter of being immune from the judgments of others, one would expect that, what constitutes the breach of privacy in the present case, would be the formation of unsolicited (and potentially negative) judgments by Sarah. But this seems implausible: After all, which new judgments could Sarah form, given that no new information has become available? If, in the Voyeuristic Sarah case, the informational model fails, as Johnson argues, then so does his immunity model, and for the same reason: where all information has been properly taken into account, there is simply no room for new judgments. By contrast, on the disattendability/civil inattention model I am proposing, Sarah's voyeurism is a breach of privacy, not because of any new information she gains or any judgments she forms, but in virtue of her paying undue attention to a routinized aspect of her target's everyday life: although her target has done nothing to attract her attention, his reasonable expectation *not to be attended to* as he goes about his business is being violated.

Similar considerations apply to the Bugged phone case: Engaging in everyday phone conversations is a routine part of modern communication; when I speak to my secret lover on the phone, both of us are operating in a way that neither intrudes upon others nor, all else being equal, should attract the attention of a third party. Indeed, enabling two – and only two – interlocutors to engage in conversations without *intruding on*, or *being intruded upon by others*, is arguably part of the proper function of telephony as a technology. (Things would be quite different if both of us were to engage in intimate conversation by shouting from the rooftops!) For a third party to intercept such a conversation would be to pay undue attention to what is, in essence, a properly disattendable activity on our part.

Finally, in the Golf on Sunday scenario, what constitutes the invasion of privacy is not the neighbour's acquisition of information about me *per se*, nor the fact that she forms a negative judgment concerning my lack of spirituality (which she would form in any case, even if she had chosen not to confront me). Rather, according to the disattendability/civil inattention model, it is the obtrusiveness of her intervention – the fact that she *lectures me on matters of morality*, when I have done nothing out of the ordinary – that renders it a violation of my privacy. (Note that, by contrast, it would hardly constitute a violation of privacy if she had volunteered, say, an unsolicited but unobtrusive comment about the weather – 'looks like a fine day for golfing' – small talk being an example of a culturally specific way of acknowledging one's neighbour without violating the norm of civil inattention.)

In other respects, too, the disattendability/civil inattention model is explanatorily superior to its rivals – the explanandum, of course, being the various culturally specific ways in which human societies allow their members to negotiate their social interactions with one another. First, it accounts for a class of cases which, as Johnson himself realizes, 'seem to present great difficulty for the judgment-of-others [=*immunity*] model'. For example,

> in spite of the well-known variation in sexual practices around the world, anthropological data indicates that sexual matters, in the main, belong to the private sphere. Societies have rules for concealment of the genitals, and restrictions on the time and manner of genital exposure; only a handful of societies practice complete nudity. (Schneider 1977: 57)

This poses a problem for the immunity model: 'The problem here is that our concern with nudity, sexuality, and other cultural taboos does not seem to be a desire for immunity from the judgment of others, as much as, the simple desire *not to be observed.*' (Johnson 1989: 165; italics added.) In defence of his immunity model, Johnson argues that such cases still involve 'a concern with judgment, rather than mere observation', and he points to the use of sexual taboos in the language we use and the 'talk of shame and embarrassment' (ibid.) in this regard. The connection between linguistic taboos and our desire for privacy in sexual matters, however, is shaky at best, and it is far from clear that the latter would disappear in less judgmental societies. By contrast, the disattendability/civil inattention model takes the 'simple desire not to be observed' as basic: Rather than treating our desire, in intimate situations, *not to be attended to* by third parties who have no legitimate claim on our attention (and vice versa) as standing in need of explanation, I propose that it be seen as a core expression of the twin principles of disattendability and civil inattention.

The disattendability/civil inattention model also holds out the promise of accounting for the historical contingency and cultural diversity of privacy as a social practice, while at the same time identifying a common underlying mechanism. In an article on Japanese conceptions of privacy, Masahiko Mizutani, James Dorsey and James Moor point out that traditional living arrangements make the spread of personal information virtually inevitable. Japanese homes are

> separated into rooms by sliding lattice-work doors (called 'shooji'), often covered with nothing more than paper. Because sound travels through these doors quite easily and as Japanese homes are rather small, there is almost surely someone in an adjoining room most of the time. In spite of this, convention has long held that conversations overheard through closed *shooji* are not to be repeated or acknowledged in any way. One can *amae* (presume indulgence) from anyone who might overhear the conversation; those who do hear will exercise *enryo*, restraint, and not act on or repeat what was overheard. (Mizutani, Dorsey, and Moor 2004: 124)

In such living conditions, it is virtually impossible to create conditions 'of not having undocumented personal knowledge about one possessed by others' (Parent 1983: 269), as the informational model would demand, nor does it seem we can suppress judgments of others, which they might privately form about us. What we can demand, though, is that we be treated with a certain level of civil inattention. (While the concept of civil inattention was initially proposed for the public realm, there is no difficulty to conceive of it as coming in degrees and as extending also to, say, individuals living under the same roof, where it would serve to give each other some 'private space'.) This appears to be exactly the function of the Japanese concept of *enryo*, restraint, which, as Mizutani et al. note, 'does much to maintain the privacy of individuals within groups' and is accompanied by an '"as if" convention for protecting privacy' (Mizutani, Dorsey, and Moor 2004: 126). It should count in favour of the disattendability/civil inattention model of privacy that it can explain the functional point of conventions such as *enryo*, which emerge even in situations that are empirically not amenable to either the containment of information *per se* (as the informational model would require) or the prevention of other people's judgments (as the immunity model would demand).

5. The Prospects of Privacy: Concluding Remarks

In this paper, I have argued for an alternative to traditional models of privacy, which either have tended to emphasize access to (or control over) information *per se*, or have aimed at preventing the formation of (potentially negative) judgments by others. By contrast, the disattendability/civil inattention model I am proposing emphasizes the *relational* aspect of privacy, which is reflected in our reasonable expectation to be accorded civil inattention by others as we go about our daily lives, without drawing undue attention to ourselves. A *state of privacy*, on this account, is a state of affairs where disattendability on the part of an individual is matched by the civil inattention the social environment accords him or her. This need not be limited to how an individual can expect to be treated in a *public* setting, but also extends to other aspects in life. For example, in the workplace we can reasonably expect that colleagues refrain from attending to, say, our love life, political orientation, or family issues. Which degree of attention is legitimate is clearly context-dependent and domain-specific, and the proposed model can accommodate such variation across contexts and domains.

A nagging worry remains, however, and stems from the earlier criticism of disattendability as an oppressive means of enforcing (often dubious) standards of 'respectability', 'civility', or the like. (See Section 4.) By linking privacy to civil inattention which, in turn, has to be 'earned' through exhibiting disattendability, the proposed account might seem to encourage the 'hiding' of aspects of an individual's behaviour, or way of life, which are merely *deemed* unacceptable by the majority, when in fact privacy should serve to protect individuals from unwarranted intrusion by others. If this worry sounds rather abstract, consider the following example. In many socially conservative societies, a gay couple holding hands or exchanging public displays of affection would no doubt attract attention and might be frowned upon; by definition, such behaviour would thereby not conform to the norm of disattendability. Yet, most proponents of a right to privacy (this author included) would deem such public disapprobation unjustified. Does the disattendability/civil inattention model then merely serve to enshrine existing (and potentially oppressive) social norms, by declaring any behaviour that might attract a significant degree of attention as 'fair game' and not worthy of privacy protection? Not quite. This is because the proposed model, while aiming to clarify the concept of privacy, is largely explanatory in character; as such, it does not pass judgment on which actions or behaviours are, or aren't, deserving of privacy. Clearly, we must leave open the possibility that there are a great number of things that *ought not to attract attention*, even if, in fact, they do. The existence of double standards for public displays of affection between heterosexual and same-sex couples is a good case in point. In short, not every violation of disattendability is a moral violation, and not every suspension of civil inattention is justified. (It is important to note that this works both ways: as we know only too well, some moral outrages – such as child abuse – are often allowed to persist because those in the know wrongly perceive them to be a 'private' matter.)

A number of authors have pointed out that privacy is an essentially cultural notion. As Johnson aptly puts it:

> What is considered private is socially or culturally defined. It varies from context to context, it is dynamic, and it is quite possible that no single example can be found of something which is considered private in every culture. (Johnson 1989: 157)

The proposed disattendability/civil inattention model acknowledges this contingency of our privacy conventions, while at the same time identifying a common mechanism. It is flexible enough to accommodate both cultural variation and technological change, insofar as *gratuitous (unwarranted) attending* may take many forms (including, say, the algorithimic monitoring of electronic communications). But the model should not be misunderstood as providing a moral justification for conformism, let alone for the suppression of minority viewpoints and uncommon ways of life, in the name of maintaining ‘disattendability’. On the contrary, it brings into sharper focus the need for a vigorous defence of the ‘visibility’ of minority practices and behaviour, given that a defence on *the grounds of privacy alone* is unlikely to be sufficiently robust and might even run the risk of downplaying the significance of public exclusion or of trivializing the much larger moral struggles to be won.

Acknowledgments. I am grateful to the participants of the Episteme conference ‘The Epistemology of Privacy and Secrecy’, held at TU Delft in May 2012, and of the NUS-Kyoto Applied Philosophy Workshop in February 2014, especially Masahiko Mizutani, for useful discussions. I would also like to thank Luke O’Sullivan (Singapore) and two anonymous reviewers for their comments on an earlier draft of this paper.

References

Martijn Blaauw. (2013) “The Epistemic Account of Privacy”, *Episteme* 19 (2), pp. 167-177.

Joshua Cohen. (2009) *Philosophy, Politics, Democracy: Selected Essays*. Cambridge, Mass.: Harvard University Press.

Don Fallis. (2013) “Privacy and Lack of Knowledge”, *Episteme* 19 (2), pp. 153-166.

Charles Fried. (1984) “Privacy [a moral analysis]”, in *Philosophical Dimensions of Privacy: An Anthology* in Ferdinand D. Schoeman (ed.), Cambridge: Cambridge University Press, pp. 203-222.

Axel Gelfert. (2006) “Kant on Testimony”, *British Journal for the History of Philosophy* 14 (4), pp. 627-652.

Raymond Geuss. (2001) *Public Good, Private Goods*. Princeton: Princeton University Press.

Erving Goffman. (1963) *Stigma: Notes on the Management of Spoiled Identity*. Englewood Cliffs: Prentice-Hall.

Erving Goffman. (1971) *Relations in Public: Microstudies of the Public Order*. New York: Basic Books.

Jeffery L. Johnson. (1989) “Privacy and the Judgment of Others”, *The Journal of Value Inquiry* 23, pp. 157-168.

Klemens Kappel. (2013) “Epistemological Dimensions of Informational Privacy”, *Episteme* 10 (2), pp. 179-192.

David Matheson. (2007) “Unknowability and Information Privacy”, *Journal of Philosophical Research* 32, pp. 251-167.

John Stuart Mill. (1963-1991) *The Collected Works of John Stuart Mill* in John M. Robson (ed.), Toronto: University of Toronto Press.

William Ian Miller. (1997) *The Anatomy of Disgust*. Cambridge, Mass.:

Harvard University Press.

Masahiko Mizutani. (2012) “Privacy, Ethics of” in *Encyclopedia of Applied Ethics* (Second Edition) in Ruth Chadwick (ed.), San Diego: Academic Press, pp. 609-615.

Masahiko Mizutani, James Dorsey, and James H. Moor. (2004), “The Internet and Japanese Conception of Privacy”, *Ethics and Information Technology* 6 (2), pp. 121-128.

W. A. Parent. (1983) “Privacy, Morality, and the Law”, *Philosophy & Public Affairs* 12 (4), pp. 269-288.

Robert C. Post. (1991) “Rereading Warren and Brandeis: Privacy, Property, and Appropriation”, *Case Western Reserve Law Review* 41, pp. 647-680.

Carl Schneider. (1977) *Shame, Exposure, and Privacy*. Boston: Beacon Press.

Daniel J. Solove. (2008) *The Future of Reputation. Gossip, Rumor, and Privacy on the Interne*t. New Haven: Yale University Press.

Judith Jarvis Thomson. (1975) “The Right to Privacy”, *Philosophy and Public Affairs* 4 (4), pp, 295-314.

Raymond Wacks. (2010) *Privacy: A Very Short Introduction*. Oxford: Oxford University Press.

Margaret Urban Walker. (2006) *Moral Repair: Reconstructing Moral Relations after Wrongdoing*, Cambridge: Cambridge University Press.

Samuel D. Warren and Louis D. Brandeis. (1890) “The Right to Privacy”, *Harvard Law Review* 193 (4), pp. 193-220.

Department of Philosophy, National University of Singapore
phigah@nus.edu.sg

철학적 분석 31 (2014) pp.183-202

Priest's Problem and Non-Injective Pluralism

Roy T Cook

Stephen Read has recently formulated an objection ('Priest's Problem') to logical pluralism (building on a thought by Graham Priest - hence the name) that suggests that one logic (the strongest of the acceptable candidates for correctness) will always be better than all others, since it is more informative with regard to the central task of logic - determining how we ought to (deductively) reason. In this paper I distinguish between two distinct forms of logical pluralism - injective pluralism (the sort defended by JC Beall and Greg Restall), and non-injective pluralism - and I then show that Priest's Problem only affects injective pluralism: non-injective pluralism is immune to the objection.

Keywords Logic, Pluralism, Intuitionism, Consequence, Reasoning

1 Monism, Pluralism, and Nihilism

It is likely that there are as many distinct versions of logical pluralism as there are logical pluralists.[1] Here we shall be interested in two general types of pluralism – what I shall call *injective* and *non-injective pluralism*. Before we can provide precise characterizations of these to general types (a task undertaken in §2), however, we need to say a bit about what a logic is, and what logics are for.

A logic is a formal codification, or model (in roughly the scientific sense of this term) of the notion of logical consequence. The classic statement regarding the nature of logical consequence is provided by Alfred Tarski, who asks us to:

> Consider any class Δ of sentences and a sentence Φ which follows from the sentences of this class. From an intuitive standpoint it can never happen that both the class Δ consists only of true sentences and the sentence Φ is false. Moreover, since we are concerned here with the concept of logical, i.e. *formal*, consequence, and thus with a relation which is to be uniquely determined by the form of the sentences between which it holds ... the consequence relation cannot be affected by replacing the designations of the objects referred to in these sentences by the designations of any other objects. ((Tarski 1983): 414 – 415)

Although there is much philosophical debate regarding how, exactly, we ought to understand the notion of logical consequence, and regarding how we ought to understand Tarski's particular take on this notion, we can isolate two crucial ideas regarding the notion which remain relatively uncontroversial – the idea that the consequence relation in natural language involves both *necessity* and *formality*. We can sum up this insight as follows:

> *LC*: A (natural language) statement Φ is a logical consequence of a set of (natural language) statements Δ if and only if:

[1] For a useful survey of various types of logical pluralism, see (Cook 2010).

> (NECESSITY) The simultaneous truth of every member of Δ guarantees the truth of Φ

and:

> (FORMALITY) This guarantee follows solely from the logical form of Φ and of the members of Δ.

If a logic is meant to be a codification of the logical consequence relation in natural language, then we require some criterion for judging when a particular formal logic (which is, after all, merely an algebraic structure) adequately or correctly captures the informal consequence relation that is the target of the codification. In short, we need to know when a logic gets it 'right', in whatever sense of 'right' turns out to be relevant. The following 'recipe' provides just such a criterion:[2]

> Given a logic $\langle \mathcal{L}, \Rightarrow \rangle$, and letting $\rhd$ represent the natural language consequence relation:
>
> 1. Identify a subset LV of the primitive symbols of $\mathcal{L}$ – this is the *logical vocabulary* of $\mathcal{L}$.
> 2. Construct a (partial) translation function T from LV to appropriate bits of natural language – this projects the logical/nonlogical distinction in $\langle \mathcal{L}, \Rightarrow \rangle$ onto our natural language.
> 3. Determine whether or not the following CORRECTNESS PRINCIPLE holds:
>
>   CP: Given any recursive mapping I from $\mathcal{L}$ to statements in our natural language (i.e. an interpretation) which agrees with T on LV, and given any statement Φ and set of statements Δ from $\mathcal{L}$:
>
> $$I(\Delta) \rhd I(\Phi) \text{ if and only if } \Delta \Rightarrow \Phi.$$

A logic is correct if and only if it satisfies the correctness principle (relative to whatever logical/non-logical distinction is privileged or itself correct). Of course, there are a number of issues raised by this brief description of the criteria by which we judge a logic to be correct that deserve further attention, including:

[2] A version of this recipe first appears in a discussion of the intuitionistic-versus-classical logic debate in (Cook 2005), and is also applied in a preliminary exploration of what we are here calling injective logical pluralism in (Cook 2014).

- How, exactly, do we draw the distinction between logical and non-logical vocabulary?[3]
- How, exactly, are we to square the centrality of model theory within logical theorizing with its absence in the description of the recipe above?[4]

Assuming that we agree on the location of the logical/non-logical divide, on how to project this formal distinction onto natural language, and on the details regarding what counts as a legitimate interpretation (assumptions that we shall make from here on), there are (at least at first glance) three possible positions that one can take with regard to the number of logics that satisfy the correctness principle. First, there is the traditional, default view – Logical Monism – the view that there is exactly *one* logic that satisfies CP:[5]

$$LM : (\exists! \Rightarrow)(\forall\Delta)(\forall\Phi)[(I(\Delta) \rhd I(\Phi)) \leftrightarrow (\Delta \Rightarrow \Phi)]$$

Second in popularity is Logical Pluralism – the view that there is *more than one logic* that satisfies CP:[6]

$$\begin{aligned} LP : (\exists \Rightarrow_1)(\exists \Rightarrow_2)[&(\exists\Delta)(\exists\Phi)((\Delta \Rightarrow_1 \Phi) \wedge (\Delta \not\Rightarrow_2 \Phi))\wedge \\ &(\forall\Delta)(\forall\Phi)((I(\Delta) \rhd I(\Phi)) \leftrightarrow (\Delta \Rightarrow \Phi))\wedge \\ &(\forall\Delta)(\forall\Phi)((I(\Delta) \rhd I(\Phi)) \leftrightarrow (\Delta \Rightarrow \Phi))] \end{aligned}$$

The most well known, and well-developed, version of Logical Pluralism is that defended in (Beall & Restall 2006), although as we shall see, other variants exist. Finally (and, demographically speaking, the least popular position) is Logical Nihilism – the view that there is *no* logic that satisfies CP, that is:

$$LN : \neg(\exists \Rightarrow)(\forall\Delta)(\forall\Phi)[(I(\Delta) \rhd I(\Phi)) \leftrightarrow (\Delta \Rightarrow \Phi)]$$

[3] Interestingly, Tarski himself might have been a pluralist of another sort, with regard to the number of legitimate ways to draw the logical/non-logical distinction. See (Varsi 2002) for details, and for a modern development of these ideas. Here we assume that the logical/non-logical distinction is unique, and fixed.

[4] My own preferred answer here is that model theory has no place in judging whether a logic is or is not correct – we do not determine whether a logic is correct via determining whether it agrees with some model theory, but rather in terms of whether it agrees with with the logical consequence relation in natural language. Model theoretic tools can play a central role in an *explanation* of why a particular logic does or does not agree with natural language logical consequence, however.

[5] Here, and below, the quantifiers binding the interpretation variable I are suppressed for readability.

[6] I am here distinguishing between *logical pluralism* – the view that there is more than one correct 'all-purpose', everywhere applicable logic – and *logical relativism* – the view that different logics are correct relative to different domains (and no logic is correct in all domains). For examples of the latter sort of view, see (Carnap 1959), (Lynch 2009), and (Pedersen 2014). Given the way we have set up the nihilism/monism/pluralism taxonomy above, logical relativism of this sort is, strictly speaking, a variant of logical nihilism (even if it obviously has philosophical affinities with pluralism).

Logical nihilism has been defended by Gillian Russell (Russell forthcoming).

It is worth noting that this way of setting up the debate – in terms of three mutually exclusive and jointly exhaustive alternatives – already presupposes that classical logic is at least one of the logics (perhaps the only one) that satisfies the correctness principle. The disjunction:

$$\textsc{Nihilism} \vee \textsc{Monism} \vee \textsc{Pluralism}$$

has roughly the following as its logical form:

$$\begin{aligned}&\neg(\exists \Rightarrow)CP(\Rightarrow) \vee (\exists! \Rightarrow)CP(\Rightarrow)\vee\\&(\exists \Rightarrow_1)(\exists \Rightarrow_2)[\Rightarrow_1 \neq \Rightarrow_2 \wedge CP(\Rightarrow_1) \wedge CP(\Rightarrow_2)]\end{aligned}$$

This formula is a logical truth in classical logic, but not in many of the non-classical logics that are in competition with the classical formalism with respect to satisfaction of the correctness principle (CP). This observation will become important in §2 and §3 below, where a version of pluralism that is (intuitionistically, but not classically) weaker than (LP) above is explored.

2 A Plurality of Pluralisms

In reality, the discussion of logical pluralism given above obscures an important complication – one that distinguishes, for example, the sort of pluralism defended in (Beall & Restall 2006), and the sort of pluralism tentatively explored in (Cook 2014).

Beall and Restall begin, as does the present essay, with Tarski's discussion of logical consequence, and from that discussion they distill the following core principle, which they call the Generalized Tarski Thesis:

> An argument is valid$_x$ if and only if, in every case$_x$ in which the premises are true, so is the conclusion. ((Beall & Restall 2006): 29).

We then, according to the Beall/Restall account, obtain different, distinct, yet equally correct logics by plugging different classes of constructions in for the relevant class of cases: if the cases are classical models, then we obtain classical logic via application of the *GTT*; if the cases are Kripke structures, then we obtain intuitionistic logic; if the cases are in inconsistent structures we obtain dialethic logic; etc.

This is certainly an interesting and theoretically substantial logical pluralism of some sort, but the details do not quite fit the codification of logical pluralism provided in §1: Beall and Restall are explicit about the fact that, on their version of logical pluralism, the different logics obtained by different choices of class of cases correctly codify *different* relations between premises and conclusions in natural language:

> We are ***pluralists*** about logical consequence because we take there to be a number of different consequence relations, each reflecting different precisifications of the pre-theoretic notion of deductive logical consequence . . . We do not take different logics to be rival analyses of the one fundamental notion (of logical consequence) because we think that the one fundamental notion of logical consequence can be made precise in different ways, each of which sharpens and disentangles the notion in different ways. These different relations are not in competition and are not rivals. ((Beall & Restall 2006): 88)

In short, according to the variant of logical pluralism defended in (Beall & Restall 2006), it is not the case that there exists more than one logic that correctly codifies the single, univocal natural language consequence relation. Rather, there are different logics, obtained by different choices of class of cases, and motivated by different emphases or concerns, that codify distinct (precise) logical consequence relations (plural!) that hold between premises and conclusions in natural language.[7]

Thus, a better way to codify the pluralism found in (Beall & Restall 2006) is something like the following, which we shall call Injective Logical Pluralism for reasons that shall be clear shortly:[8]

$$\begin{aligned} ILP: \ & (\exists \rhd_1)(\exists \rhd_2)(\exists \Rightarrow_1)(\exists \Rightarrow_2) \\ & [(\exists \Delta)(\exists \Phi)((I(\Delta) \rhd_1 I(\Phi)) \wedge (I(\Delta) \not\rhd_2 I(\Phi)))\wedge \\ & (\exists \Delta)(\exists \Phi)((\Delta \Rightarrow_1 \Phi) \wedge (\Delta \not\Rightarrow_2 \Phi))\wedge \\ & (\forall \Delta)(\forall \Phi)((I(\Delta) \rhd_1 I(\Phi)) \leftrightarrow (\Delta \Rightarrow_1 \Phi))\wedge \\ & (\forall \Delta)(\forall \Phi)((I(\Delta) \rhd_2 I(\Phi)) \leftrightarrow (\Delta \Rightarrow_2 \Phi))] \end{aligned}$$

In short, injective logical pluralism amounts to the claim that there are (at least) two distinct relations in natural language worthy of the honorific "logical consequence", and there are (at least) two distinct logics, each of which satisfies the correctness principle (CP) with respect to exactly one of these relations on natural language.

Once this observation is made, however, another way of formulating an interesting, and theoretically substantial, version of logical pluralism becomes obvious: we can take the original formulation of logical pluralism

[7] Some care should be taken here, since each of the precise logical consequence relations posited by (Beall & Restall 2006) is a precisification of the single, imprecise yet fundamental notion. Thus, (Beall & Restall 2006) do not deny that there is a single, univocal and fundamental notion of logical consequence, but instead deny (in effect) that any formal logic exactly captures this notion (rather than capturing one or the other of the many different precisifications of it).

[8] Carnap's *logical tolerance*, as explicated in (Carnap 1959), where different logics are correct with respect to distinct linguistic frameworks with their own consequence relations, is also a version of injective logical pluralism.

– that is, (LP) – at face value, and develop a view where there are distinct logics that each satisfies the correctness principle with respect to a single, univocal logical consequence relation in natural language. We can easily formulate such a view – which we shall call NON-INJECTIVE LOGICAL PLURALISM in virtue of the fact that it 'maps' more than one logic to a single consequence relation – along lines similar to the explicit statement of injective logical pluralism:[9]

$$\begin{aligned} nILP: \ & (\exists \rhd)(\exists \Rightarrow_1)(\exists \Rightarrow_2) \\ & [(\exists \Delta)(\exists \Phi)((\Delta \Rightarrow_1 \Phi) \wedge (\Delta \not\Rightarrow_2 \Phi)) \wedge \\ & (\forall \Delta)(\forall \Phi)((I(\Delta) \rhd I(\Phi)) \leftrightarrow (\Delta \Rightarrow_1 \Phi)) \wedge \\ & (\forall \Delta)(\forall \Phi)((I(\Delta) \rhd I(\Phi)) \leftrightarrow (\Delta \Rightarrow_2 \Phi))] \end{aligned}$$

At this point we might ask why Beall and Restall opted for injective, rather than non-injective, pluralism. I will not speculate on their actual reasons (since I doubt they considered the question in anything like these terms), but we can point out one reason why they were right to choose the path they in fact chose.

If one is a logical pluralist (of any sort), and if in addition one thinks that (all of) the correct (or best, or legitimate, or whatever) logics are topic neutral and completely general – that is, if one thinks that any 'correct' logic ought to be applicable to any subject matter – then these logics should apply to metaphilosophical/metalogical reasoning about the role of logic itself. As a result, one's formulation of logical pluralism must be consistent within any and all of the logics that turn out to be correct.[10] $nILP$ is inconsistent,

[9] Of course, for the view to be interesting, there must be more to it than the observation that there are distinct logics that correctly codify some relation or other on natural language – that much is trivial. Beall and Restall do supply the ingredients for such an account, by arguing that any logical consequence relation on natural language must be necessary, normative, and formal – see (Beall & Restall 2006): 14 – 23.

[10] Unless, perhaps, one is a pluralist and a dialethiest – that is, unless one believes that multiple logics are correct and that all of the correct logics allow for true contradictions. Even here, however, we can and should still require that the formulation of logical pluralism be non-trivial, even if it might be inconsistent. I will not explore this interesting option here, however.

however, as the following derivation demonstrates:

1	$(\exists\Delta)(\exists\Phi)((\Delta \Rightarrow_1 \Phi) \wedge (\Delta \not\Rightarrow_2 \Phi))$	Premise.
2	$(\forall\Delta)(\forall\Phi)((I(\Delta) \rhd I(\Phi)) \leftrightarrow (\Delta \Rightarrow_1 \Phi))$	Premise.
3	$(\forall\Delta)(\forall\Phi)((I(\Delta) \rhd I(\Phi)) \leftrightarrow (\Delta \Rightarrow_2 \Phi))$	Premise.
4	$\quad\vert\ (\Delta \Rightarrow_1 \Phi) \wedge (\Delta \not\Rightarrow_2 \Phi)$	Assumption (for $\exists E$).
5	$\quad\vert\ \Delta \Rightarrow_1 \Phi$	4.
6	$\quad\vert\ I(\Delta) \rhd I(\Phi)$	2, 5.
7	$\quad\vert\ \Delta \Rightarrow_2 \Phi$	3, 6.
8	$\quad\vert\ \Delta \not\Rightarrow_2 \Phi$	4.
9	$\quad\vert\ \bot$	7, 8.
10	$\bot$	1, 4 - 9, $\exists E$.

It is worth noting that this derivation shows that $nILP$ is not only classically inconsistent, but is also intuitionistically (in fact, minimally) inconsistent. Thus, *merely* moving to a constructive setting is of no help with the present problem. Since classical logic and intuitionistic logic are two of the logics that (Beall & Restall 2006) take to be paradigm instances of correct logics, $nILP$ is not an option for them. Furthermore, even if (unlike Beall and Restall) we were willing to give up on the correctness (even plurally) of classical logic, adopting $nILP$ entails that any correct logic is either weaker than, or incomparable to, intuitionistic logic. While developing such a pluralism about logic, where all of the correct logics are extremely weak, is perhaps not impossible, here we shall instead look for alternatives that are somewhat more generous in terms of the logics that might be candidates for correctness.

There is a move we can make: we can reformulate a version of non-injective logical pluralism that it is not inconsistent, at least in some logics stronger than intuitionistic logic. One way to do so is to replace the formulation given above with the following, which we shall call Weak non-Injective Logical Pluralism:

$$\begin{aligned} WnILP: \ & (\exists\rhd)(\exists \Rightarrow_1)(\exists \Rightarrow_2) \\ & [\neg(\forall\Delta)(\forall\Phi)((\Delta \Rightarrow_1 \Phi) \leftrightarrow (\Delta \Rightarrow_2 \Phi))\wedge \\ & (\forall\Delta)(\forall\Phi)(\neg(I(\Delta) \rhd I(\Phi)) \leftrightarrow \neg(\Delta \Rightarrow_1 \Phi))\wedge \\ & (\forall\Delta)(\forall\Phi)(\neg(I(\Delta) \rhd I(\Phi)) \leftrightarrow \neg(\Delta \Rightarrow_2 \Phi))] \end{aligned}$$

Weak non-injective logical pluralism is, of course, classically inconsistent, since it is classically equivalent to $nILP$. But $WnILP$ is intuitionistically rather weaker than $nILP$, and is consistent in intuitionistic logic, as the following result demonstrates:[11]

Theorem 2.1. *There is a Kripke structure $\mathcal{M}$ such that:*

$$\mathcal{M} \models WnILP$$

Proof. $\mathcal{M}$ is as follows:

$$\begin{aligned} \mathcal{M} &= <W, R, I> \\ W &= \{w_n : n \in \mathbb{N}\} \qquad \text{(Domain of each } w_n = \mathbb{N}) \\ I(\Rightarrow_1, w_n) &= \{\langle m, m+1\rangle : m \geq n\} \\ I(\Rightarrow_2, w_n) &= \{\langle m, m+1\rangle : m > n\} \\ I(\rhd_1, w_n) &= \{\langle m, m+1\rangle : m > n+1\} \end{aligned}$$

□

In fact, weak non-injective logical pluralism is consistent in logics much stronger than intuitionistic logic (even if not in classical logic). For example, $WnILP$ is consistent in Gödel-Dummett logic (LC), which we obtain by adding the linearity axiom:

$$(\Phi \rightarrow \Psi) \vee (\Psi \rightarrow \Phi)$$

to intuitionistic logic (H), and is consistent in any of the (continuum-many) logics between LC and H, as the following corollary demonstrates:

Corollary 2.2. *$WnILP$ is consistent in any logic L such that $H \subseteq L \subseteq LC$.*

Proof. First order LC is sound and complete with respect to the class of linearly-order Kripke structures, see (Corsi 1992). The model constructed in the proof of Theorem 2.1 is linearly ordered. Hence $WnILP$ is consistent in LC, and in any logic strictly weaker than LC. □

This suggests an interesting non-injective version of logical pluralism where one accepts as correct only logics within which $WnILP$ is consistent. Of course, we need not accept all such logics, and in (Cook 2014) I explore a version of pluralism that privileges those logics within which $WnILP$ is consistent, and which are also in good standing with respect to the semantic, proof-theoretic, and metaphysical challenges proposed by constructivists (e.g. those challenges found in (Heyting 1974), (Dummett

[11] The model constructed in the proof involves a simplification, since it represents both formulas (conclusions) and sets of formulas (premise sets) as natural numbers. We can understand this proof as handling the single-premise case directly, however, with the multiple-premise generalization following from the fact that intuitionistic logic is compact with respect to Kripke structures.

1991), and (Tennant 1996)).[12] The details regarding exactly which logics survive these challenges are complex – I refer the reader to the relevant sections of (Cook 2014) for details – but here we can focus on a slightly simpler position: acceptance of $WnILP$ where the correct logics are exactly those logics lying between LC and H. All of the points made here will also apply to the slightly different collection of logics examined in the earlier paper.

I will not attempt to provide a positive argument here for the claim that this particular version of pluralism is correct – that is another task for another day. Instead, we shall look at a particular criticism of logical pluralism, and I shall demonstrate that the objection only applies to injective logical pluralism, and not to non-injective logical pluralism (at least, not to the version that takes the correct logics to lie roughly between H and LC).

3 Priest's Problem

Now that we have our two competing variants of pluralism carefully formulated, it is worth exploring how they fare with respect to extant criticisms of logical pluralism. In particular, we will examine an objection to logical pluralism attributed to Graham Priest by Stephen Read.[13]

The objection in question begins by noting that the primary reason we are interested in logics, and are interested in which logic (or logics) is (or are) correct, legitimate, or best is that logics are meant to codify correct reasoning. In particular, if Φ and Δ are a formula and set of formulas in some formal language, and I is an appropriate interpretation function mapping formulas to sentences of natural language, then if the image of Φ under I is a logical consequence of the image of Δ under I, and the image

[12] The particular constraints I take to follow from these challenges are that the logic in question must:

1. Validate the standard introduction and elimination rules (i.e. be at least as strong as either minimal or intuitionistic logic, depending on the details).
2. Have the disjunction property (or be contained in a logic with the disjunction property).
3. Fails to be finitely bounded (i.e. there is no finite n and class of frames K such that the logic is sound and complete with respect to K and each frame in K has no more than n nodes.

For details, see (Cook 2014).

[13] The objection reconstructed in (Read 2006) and called "Priest's Challenge" is not exactly the objection actually formulated in (Priest 2001). The point intended by Priest – that, when working within the framework of (Beall & Restall 2006) and confronted by two logics based on two overlapping classes of cases, we should reason in accordance with the logic that results from considering the intersection of the two overlapping classes – is distinct from (although intimately connected to) the problem discussed by Read, and is (unlike Read's variant) specific to the pluralism developed by Beall and Restall. Since the worry identified by Read is a genuine worry, and since he attributes it to Priest, I shall continue to call it Priest's Problem here.

under I of each formula in Δ is true, then we are guaranteed that the image of Φ is true as well. Using the notation introduced earlier, we can codify this as the following TRUTH PRINCIPLE:

$$TP : (\forall\Delta)(\forall\Phi)(((I(\Delta) \rhd I(\Phi)) \wedge (\forall\Psi \in \Delta)(I(\Psi))) \rightarrow I(\Phi))$$

Even more importantly, there is a normative component to the undertaking in question: if the image of Φ under I is a logical consequence of the image of Δ under I, and if all we believe all of the sentences in the image of Δ (equivalently: we believe the images, under I, of each formula in Δ), then (insofar as we take a doxastic stand on $I(\Phi)$ in the first place) we *ought* to believe $I(\Phi)$. Priest's Problem now proceeds as follows (see (Read 2006): 194 – 196 for the original informal version of the argument)[14]: Let $\langle\mathcal{L}, \Rightarrow_1\rangle$ and $\langle\mathcal{L}, \Rightarrow_2\rangle$ be two distinct, 'correct' logics (on a shared formal language $\mathcal{L}$) where we understand correctness in terms of the injective logical pluralism account, and let Φ and Δ be a formula and set of formulas respectively such that $\Delta \Rightarrow_1 \Phi$ and $\Delta \not\Rightarrow_2 \Phi$. Now, assume that I is some appropriate interpretation function mapping the formulas of our logic to natural language, and assume that all of the sentences in the image of Δ under I are true – that is, assume that:

$$(\forall\Psi \in \Delta)(I(\Psi))$$

[14] (Read 2006) takes things a step further, considering two logics $\langle\mathcal{L}, \Rightarrow_1\rangle$ and $\langle\mathcal{L}, \Rightarrow_2\rangle$ where there is a formula Φ and set of formulas Δ such that:

$$\Delta \Rightarrow_1 \Phi$$
$$\Delta \Rightarrow_2 \neg\Phi$$

The particular case he considers involves classical logic (C) and Abelian logic (A) – where:

$$\neg A, B \Rightarrow_C \neg((((A \rightarrow B) \rightarrow B) \rightarrow A))$$
$$\neg A, B \Rightarrow_A ((((A \rightarrow B) \rightarrow B) \rightarrow A))$$

While such situations are of interest to particularly permissive formulations of logical pluralism, they are orthogonal to the present concern, since weak non-injective logical pluralism restricts attention to logics between intuitionistic and classical (in fact, to those between intuitionistic and Gödel-Dummett), and within such a range of logics no such conflicts can arise.

We can now reason as follows (Δ and Φ fixed):

1	$\Delta \Rightarrow_1 \Phi$	Premise.
2	$(I(\Delta) \rhd_1 I(\Phi)) \leftrightarrow (\Delta \Rightarrow_1 \Phi)$	Premise.
3	$((I(\Delta) \rhd_1 I(\Phi)) \wedge (\forall \Psi \in \Delta)(I(\Psi))) \rightarrow I(\Phi)$	Premise.
4	$(\forall \Psi \in \Delta)(I(\Psi))$	Premise
5	$I(\Delta) \rhd_1 I(\Phi)$	1, 2.
6	$(I(\Delta) \rhd_1 I(\Phi)) \wedge (\forall \Psi \in \Delta)(I(\Psi)))$	4, 5.
7	$I(\Phi)$	3, 6.

In other words, if all of the natural language sentences corresponding to the members of Δ are true, then $\langle \mathcal{L}, \Rightarrow_1 \rangle$ tells us that the natural language sentence corresponding to Φ is true. In the same situation, however, $\langle \mathcal{L}, \Rightarrow_2 \rangle$ tells us nothing about whether or not the natural language sentence corresponding to Φ is true or false.[15] Working with the analogous premises with respect to $\langle \mathcal{L}, \Rightarrow_2 \rangle$:

1	$\Delta \not\Rightarrow_2 \Phi$	Premise.
2	$(I(\Delta) \rhd_2 I(\Phi)) \leftrightarrow (\Delta \Rightarrow_2 \Phi)$	Premise.
3	$((I(\Delta) \rhd_2 I(\Phi)) \wedge (\forall \Psi \in \Delta)(I(\Psi))) \rightarrow I(\Phi)$	Premise.
4	$(\forall \Psi \in \Delta)(I(\Psi))$	Premise

(which differ from the premises used in the derivation just given solely in virtue of the substitution of $\Rightarrow_2$ for $\Rightarrow_1$ and the presence of a negation in the first premise), we cannot prove that $I(\Phi)$ is true, nor can we prove that it is false. In fact, we can prove nothing relevant about the semantic status of $I(\Phi)$ at all: $\langle \mathcal{L}, \Rightarrow_2 \rangle$ is completely silent with regard to the status of $I(\Phi)$. Read sums up the issue as follows:

> It follows that in a very real sense $\langle \mathcal{L}, \Rightarrow_1 \rangle$ and $\langle \mathcal{L}, \Rightarrow_2 \rangle$ are not equally good. $\langle \mathcal{L}, \Rightarrow_1 \rangle$ answers a crucial question that $\langle \mathcal{L}, \Rightarrow_2 \rangle$ does not. ((Read 2006): 195, notation altered to match that used here.)

In short, whenever two logics disagree with regard to a particular argument (form) – one validating it and the other not – and the natural language

[15] Actually, this is not quite right: $\langle \mathcal{L}, \Rightarrow_2 \rangle$, in some indirect sense, at least, 'tells' us that we should not conclude that $I(\Phi)$ is true based solely on purely logical facts and the fact that $I(\Psi)$ is true for all $\Psi \in \Delta$. This does not affect the point, however, which is that $\langle \mathcal{L}, \Rightarrow_2 \rangle$ provides no information *simpliciter* regarding the truth value (or lack thereof, for that matter) of $I(\Phi)$.

sentences corresponding to the premises of the argument are true, we should reason according to the (locally) stronger logic, and infer that the conclusion of the argument is true as well. It is a small step from this observation to a complete collapse of injective logical pluralsim, since it seems that we should therefore reason, in all situations, in accordance with the logic obtained by 'unioning' up all of the 'correct' logics – that is, the truly best logic is the 'super'-logic $\langle \mathcal{L}, \Rightarrow^{+} \rangle$ where:

$$(\forall\Delta)(\forall\Phi)((\Delta \Rightarrow^{+} \Phi) \leftrightarrow (\exists \Rightarrow_{\alpha})((\Delta \Rightarrow_{\alpha} \Phi) \wedge \\ (\exists \rhd_{\alpha})(\forall\Delta_2)(\forall\Phi_2)(I(\Delta_2) \rhd_{\alpha} I(\Phi_2) \leftrightarrow (\Delta \Rightarrow_{\alpha} \Phi))))$$

In short, an argument is valid in the 'super'-logic $\langle \mathcal{L}, \Rightarrow^{+} \rangle$ if and only if it is valid in some logic that satisfies the correctness principle with respect to some legitimate natural language consequence relation. But if this logic is the right one to apply in all cases, then there seems little reason to accept pluralism (at least, injective logical pluralism) with its plethora of weaker, less informative logics.

Priest's Problem is thus a genuine problem for injective logical pluralsim. But is it a problem for weak non-injective logical pluralism? The answer, surprisingly, is "no": let $\langle \mathcal{L}, \Rightarrow_1 \rangle$ and $\langle \mathcal{L}, \Rightarrow_2 \rangle$ be two distinct, 'correct' logics (on a shared formal language $\mathcal{L}$) where we understand correctness in terms of the weak non-injective logical pluralism account (hence, $\Rightarrow_1$ and $\Rightarrow_2$ lie between H and LC) and let Φ and Δ be a formula and set of formulas respectively such that $\Delta \Rightarrow_1 \Phi$ and $\Delta \not\Rightarrow_2 \Phi$.[16] Now, assume that I is some appropriate interpretation function mapping the formulas of our logic to natural language, and assume that all of the sentences in the image of Δ under I are true – that is, assume that:

$$(\forall\Psi \in \Delta)(I(\Psi))$$

We cannot reason exactly as we did when considering injective logical pluralism, since the formulation of correctness (Premise 2 below) is substantially weaker (from the perspective of any logic between H and LC) in weak non-injective logical pluralism. We can, however, prove, not that the image of Φ

[16] Note that this assumption is stronger than what is minimally required by $WnILP$, since $WnILP$ only requires that the logics fail to agree everywhere:

$$\neg(\forall\Delta)(\forall\Phi)((\Delta \Rightarrow_1 \Phi) \leftrightarrow (\Delta \Rightarrow_2 \Phi))$$

and not that they necessarily disagree on some *particular* inference:

$$\exists(\Delta)(\exists\Phi)((\Delta \Rightarrow_1 \Phi) \wedge (\Delta \not\Rightarrow_2 \Phi))$$

Using the stronger, existential formulation in the argument above simplifies the reasoning, however.

under I is true, but rather that the double negation of $I(\Phi)$ is true:

1	$\Delta \Rightarrow_1 \Phi$	Premise.
2	$\neg(I(\Delta) \rhd I(\Phi)) \leftrightarrow \neg(\Delta \Rightarrow_1 \Phi)$	Premise.
3	$((I(\Delta) \rhd I(\Phi)) \wedge (\forall\Psi \in \Delta)(I(\Psi))) \rightarrow I(\Phi)$	Premise.
4	$(\forall\Psi \in \Delta)(I(\Psi))$	Premise
5	$\neg\neg(I(\Delta) \rhd I(\Phi))$	1, 2.
6	$\neg\neg((I(\Delta) \rhd I(\Phi)) \wedge (\forall\Psi \in \Delta)(I(\Psi)))$	4, 5.
7	$\neg\neg I(\Phi)$	3, 6.

Thus, according to weak non-injective logical pluralism, if a logic 'tells' us that some conclusion follows from a set of premises, and if the premises are all true, then we need not (in fact, should not!) infer that the conclusion is true, but only that its double negation is.[17]

At this point the reader might be forgiven for suspecting that we haven't made much progress, since the problem with injective logical pluralism wasn't specifically that one of the logics allowed us to infer the *truth* (rather than the truth of the double negation) of the conclusion, but resulted from the fact that the first logic allowed us to prove something – anything – semantically informative about the conclusion while the second did not. Here, however, we can take advantage of the fact that weak non-injective logical pluralism does not allow just any logic to count as legitimate or correct. On the contrary, it was suggested above that the only plausible candidates for correct logics, on this view, lie between intuitionistic logic and Gödel-Dummett logic. The following theorem is a straightforward consequence of well-known facts holding of logics between classical and intuitionsitic:

Theorem 3.1. *For any propositional logics $\langle \mathcal{L}, \Rightarrow_1 \rangle$ and $\langle \mathcal{L}, \Rightarrow_2 \rangle$ such that:*

$$H \subseteq \langle \mathcal{L}, \Rightarrow_1 \rangle \subseteq LC$$
$$H \subseteq \langle \mathcal{L}, \Rightarrow_2 \rangle \subseteq LC$$

if:

$$\Delta \Rightarrow_1 \Phi$$

then:

$$\Delta \Rightarrow_2 \neg\neg\Phi$$

[17] An important clarification: This strange conclusion is not equivalent to the claim that if the sentences in Δ are true (in a model/interpretation/etc.) then the conclusion Φ need not be true, but only $\neg\neg\Phi$. Rather, the initially strange-appearing double negation claim, properly understood, concerns the not-quite-perfect 'fit' between formal language and natural language.

Proof. Assume $\Delta \Rightarrow_1 \Phi$. Then, since $\langle \mathcal{L}, \Rightarrow_1 \rangle \subseteq LC \subset C$, $\Delta \Rightarrow_C \Phi$. The following is a well-known fact about intuitionistic and classical propositional logic (see (Troelstra & van Dalen 1988) for details):

$$(\forall \Delta)(\forall \Phi)((\Delta \Rightarrow_C \Phi) \leftrightarrow (\Delta \Rightarrow_H \neg\neg\Phi))$$

Hence, $\Delta \Rightarrow_H \neg\neg\Phi$. Then, since $H \subseteq \langle \mathcal{L}, \Rightarrow_2 \rangle$, $\Delta \Rightarrow_2 \neg\neg\Phi$. □

Given this fact, we can now reason as follows with respect to $\langle \mathcal{L}, \Rightarrow_2 \rangle$:[18]

1	$\Delta \not\Rightarrow_2 \Phi$	Premise.
2	$\Delta \Rightarrow_1 \Phi$	Premise
3	$\neg(I(\Delta) \rhd I(\neg\neg\Phi)) \leftrightarrow \neg(\Delta \Rightarrow_2 \neg\neg\Phi)$	Premise.
4	$((I(\Delta) \rhd I(\neg\neg\Phi)) \wedge (\forall \Psi \in \Delta)(I(\Psi))) \rightarrow I(\neg\neg\Phi)$	Premise.
5	$(\forall \Psi \in \Delta)(I(\Psi))$	Premise
6	$\Delta \Rightarrow_2 \neg\neg\Phi$	2, Theorem 3.1.
7	$\neg\neg(I(\Delta) \rhd I(\neg\neg\Phi))$	1, 2.
8	$\neg\neg((I(\Delta) \rhd I(\neg\neg\Phi)) \wedge (\forall \Psi \in \Delta)(I(\neg\neg\Psi)))$	4, 5.
9	$\neg\neg I(\neg\neg\Phi)$	3, 6.

In short, even if $\langle \mathcal{L}, \Rightarrow_2 \rangle$ does not validate the argument from Δ to Φ, it nevertheless tells us that, if all of the natural language sentences corresponding to the formulas in Δ are true, then the (natural language) double negation of the natural language sentence corresponding to $\neg\neg\Phi$ is true. Thus, unlike the case with injective logical pluralism, on the present view both logics tell us something about the semantic status of the conclusion of an argument, even if only one of the logics strictly speaking validates the relevant argument.[19]

[18] Note that we have modified the particular instances of the CORRECTNESS PRINCIPLE and the TRUTH PRINCIPLE used in the argument.

[19] Of course, the fact about double negations that played a central role in this argument – that is:

$$(\forall \Delta)(\forall \Phi)((\Delta \Rightarrow_C \Phi) \leftrightarrow (\Delta \Rightarrow_H \neg\neg\Phi))$$

does not hold when the logic is first- (or higher-) order. There are recursive translations $T(\dots)$, which involve inserting double negations within the formula in question (for example, prefixed to each existential quantifier) such that:

$$(\forall \Delta)(\forall \Phi)((\Delta \Rightarrow_C \Phi) \leftrightarrow (\Delta \Rightarrow_H T(\Phi)))$$

holds for first (and higher-) order sentences. Further, we can easily construct such a translation function so that in addition we have:

$$(\forall \Delta)(\forall \Phi)(\Delta \Rightarrow_H T(\Phi) \leftrightarrow \Delta \Rightarrow_H \neg\neg T(\Phi))$$

In addition, there are good reasons for thinking that $\langle \mathcal{L}, \Rightarrow_1 \rangle$ and $\langle \mathcal{L}, \Rightarrow_2 \rangle$ tell us *exactly the same thing* with respect to the status of Φ. If, as seems entirely reasonable, we require that any interpretation function I mapping our formal language to natural language must be such that the interpretation of a negated formula is the (natural language) negation of that formula:[20]

$$(\forall \Phi)(\neg I(\Phi) = I(\neg \Phi))$$

then it follows that $\langle \mathcal{L}, \Rightarrow_1 \rangle$ and $\langle \mathcal{L}, \Rightarrow_2 \rangle$ entail equivalent claims regarding the status of the natural language sentence $I(\Phi)$:

$$\begin{aligned} \neg\neg I(\neg\neg\Phi) &\leftrightarrow \neg I(\neg\neg\neg\Phi) \\ &\leftrightarrow \neg I(\neg\Phi) \\ &\leftrightarrow \neg\neg I(\Phi) \end{aligned}$$

The second step is justified by the fact that all logics between H and LC (and in fact all logics between H and C) agree on the logical equivalence of $\neg\Phi$ and $\neg\neg\neg\Phi$.

To sum up: Even if $\langle \mathcal{L}, \Rightarrow_1 \rangle$ and $\langle \mathcal{L}, \Rightarrow_2 \rangle$ do not validate the same inferences, according to weak non-injective logical pluralism they imply the same things regarding the semantic status of a given natural language conclusion, given the truth of an appropriately relevant set of natural language premises. Unlike the case with injective logical pluralism, there is no reason to privilege one or the other of the logics in terms of the answers they with regard to which natural language inferences we should make, and which natural language statements we should take to be true (or take to have true double-negations, etc.). Thus, Priest's Problem disappears.

4 The Prospects for (non-Injective) Pluralism

In the previous sections we identified two main approaches to logical pluralism: injective logical pluralism, where two or more distinct formal logics are correct relative to distinct natural language consequence relations (on the same natural language), and non-injective logical pluralism, where two or more distinct formal logics are correct relative to a single, univocal natural language consequence relation. After rejecting a strong version of non-injective pluralism, and opting for a weaker (but coherent) formulation

Thus, we could construct a version of the main argument of this paper for first-order logic in a straightforward, although much more complicated, manner analogous to the propositional case.

[20] At this point it is important to be clear about a harmless ambiguity in the notation used above: we have used $\neg$ for both the unary negation connective in our formal language(s) (when, e.g., the symbol is within the scope of "$I(\dots)$" or "$\dots \Rightarrow \dots$", and for the natural language word "not" (and its synonyms). As long as the reader is careful regarding the context, however, no harm comes of this.

of this idea, we then tested these two approaches against Priest's Problem, and determined that, although Priest's Problem presents a genuine difficulty for injective pluralism, non-injective pluralism seems immune to the worry in question.

Of course, this falls far short of a full defense of weak non-injective logical pluralism. In particular, there is one particular worry – hinted at in the discussion of the previous section – regarding the coherence of the view that requires more attention. The problem arises from the fact that a slightly strengthened version of $WnILP$ that results from replacing the first, negated universal, condition with the intuitionistically stronger existential condition – that is:

$$\begin{aligned} W^{*}nILP: \ & (\exists \rhd)(\exists \Rightarrow_1)(\exists \Rightarrow_2) \\ & [(\exists\Delta)(\exists\Phi)((\Delta \Rightarrow_1 \Phi) \wedge (\Delta \not\Rightarrow_2 \Phi)) \wedge \\ & (\forall\Delta)(\forall\Phi)(\neg(I(\Delta) \rhd I(\Phi)) \leftrightarrow \neg(\Delta \Rightarrow_1 \Phi)) \wedge \\ & (\forall\Delta)(\forall\Phi)(\neg(I(\Delta) \rhd I(\Phi)) \leftrightarrow \neg(\Delta \Rightarrow_2 \Phi))] \end{aligned}$$

is, like the strong version $nILP$, inconsistent, as evidenced by the following derivation:

1	$(\exists\Delta)(\exists\Phi)((\Delta \Rightarrow_1 \Phi) \wedge (\Delta \not\Rightarrow_2 \Phi))$	Premise.
2	$(\forall\Delta)(\forall\Phi)(\neg(I(\Delta) \rhd I(\Phi)) \leftrightarrow \neg(\Delta \Rightarrow_1 \Phi))$	Premise.
3	$(\forall\Delta)(\forall\Phi)(\neg(I(\Delta) \rhd I(\Phi)) \leftrightarrow \neg(\Delta \Rightarrow_2 \Phi))$	Premise.
4	$(\Delta \Rightarrow_1 \Phi) \wedge (\Delta \not\Rightarrow_2 \Phi)$	Assumption (for $\exists E$).
5	$\Delta \Rightarrow_1 \Phi$	4.
6	$\neg\neg(I(\Delta) \rhd I(\Phi))$	2, 5.
7	$\neg\neg(\Delta \Rightarrow)_2\Phi$	3, 6.
8	$\neg(\Delta \Rightarrow_2 \Phi)$	4.
9	$\perp$	7, 8.
10	$\perp$	1, 4 - 9, $\exists E$.

Thus, the coherence of the weak non-injective logical pluralist view is extramely sensitive to its formulation, and in particular it requires that, given any pair of legitimate logics $\langle \mathcal{L}, \Rightarrow_1 \rangle$ and $\langle \mathcal{L}, \Rightarrow_1 \rangle$, we can prove at most that they fail to agree everywhere, but cannot prove of a particular inference that it is valid in one logic and not in the other.

Note that this implies that the actual range of correct logics, on the weak non-injective logical pluralist view, must be significantly narrower than the class of logics between H and LC. The reason is simple: we can, in fact, prove:

$$(\exists\Delta)(\exists\Phi)((\Delta \Rightarrow_{LC} \Phi) \wedge (\Delta \not\Rightarrow_{H} \Phi))$$

since we can prove:

$$\varnothing \Rightarrow_{LC} ((A \rightarrow B) \vee (B \rightarrow A))$$
$$\varnothing \not\Rightarrow_{H} ((A \rightarrow B) \vee (B \rightarrow A))$$

Proof sketch: We can (constructively, that is, within a metatheory utilizing only intuitionistic logic) prove that H is sound with respect to the Kripke structure semantics (proving completeness is, of course, another matter!) But then, the (finite!) countermodel to $(A \rightarrow B) \vee (B \rightarrow A)$ shows that this formula is not provable within H, but it is of course an axiom of LC.

As a result, weak non-injective logical pluralism requires something like the following: There are two logics ($\langle \mathcal{L}, \Rightarrow_1 \rangle$ and $\langle \mathcal{L}, \Rightarrow_2 \rangle$ such that:

1. $\langle \mathcal{L}, \Rightarrow_2 \rangle$ is a sublogic of $\langle \mathcal{L}, \Rightarrow_1 \rangle$ – that is, we can prove (in a metatheory whose logic is $\langle \mathcal{L}, \Rightarrow_2 \rangle$):

$$(\forall\Delta)(\forall\Phi)((\Delta \Rightarrow_2 \Phi) \rightarrow (\Delta \Rightarrow_1 \Phi))$$

2. We can prove (in a metatheory whose logic is $\langle \mathcal{L}, \Rightarrow_2 \rangle$):

$$\neg(\forall\Delta)(\forall\Phi)((\Delta \Rightarrow_1 \Phi) \wedge (\Delta \Rightarrow_2 \Phi))$$

   I.e. we can prove that $\langle \mathcal{L}, \Rightarrow_2 \rangle$ is ‘weakly’ a proper sublogic of $\langle \mathcal{L}, \Rightarrow_1 \rangle$.

3. We cannot prove (in a metatheory whose logic is $\langle \mathcal{L}, \Rightarrow_1 \rangle$):

$$(\exists\Delta)(\exists\Phi)((\Delta \Rightarrow_1 \Phi) \wedge (\Delta \not\Rightarrow_2 \Phi))$$

   I.e. we cannot prove that $\langle \mathcal{L}, \Rightarrow_2 \rangle$ is ‘strongly’ a proper sublogic of $\langle \mathcal{L}, \Rightarrow_1 \rangle$.

Given two such logics $\langle \mathcal{L}, \Rightarrow_1 \rangle$ and $\langle \mathcal{L}, \Rightarrow_2 \rangle$, one could adopt weak non-injective pluralism with respect to the logics between $\langle \mathcal{L}, \Rightarrow_1 \rangle$ and $\langle \mathcal{L}, \Rightarrow_2 \rangle$. Thus, a complete defense of weak non-injective logical pluralism requires identifying such a range of logics – a task left for another day.

References

Beall, Jc, & Greg Restall (2006), *Logical Pluralism*, Oxford: Oxford University Press.

Carnap, Rudolph (1959), *The Logical Syntax of Language*, Paterson NJ: Littelfield, Adams & Co..

Cook, Roy (2005), "Intuitionism Reconsidered", in (Shapiro 2005), pp. 387-411.

Cook, Roy (2010), "Let a Thousand Flowers Bloom: A Tour of Logical Pluralism", *Philosophy Compass* 5, pp. 492-504.

Cook, Roy (2014), "Should Antirealists be Antirealists About Antirealism?", *Erkenntnis* 79(2), pp. 233-258.

Corsi, Giovanna (1992), "Completeness theorem for Dummett's LC Quantified and Some of its Extensions", *Studia Logica* 51(2), pp. 317-335.

Devidi, David & Tim Kenyon (eds.) (2006), *A Logical Approach to Philosophy*, Dordrecht: Springer.

Dummett, M. (1991) *The Logical Basis of Metaphysics*, Cambridge, MA: Harvard University Press.

Heyting, A. (1974), "Intuitionistic Views on the Nature of Mathematics", *Synthese* 27, pp. 79-91.

Lynch, Michael (2009), *Truth as One and Many*, Oxford: Oxford University Press.

Pedersen, Nikolaj (2014), "Pluralism×3: Truth, Logic, Metaphysics", *Erkenntnis* 79, pp. 259-277.

Priest, Graham (2001), "Logic, One or Many?", in (Woods & Brown 2001), pp. 23-38.

Read, Stephen (2006), "Monism, The One True Logic", in (Devidi &

Kenyon 2006), pp. 193-209.

Russell, Gillian (forthcoming) “Could There Be No Logic?”.

Shapiro, Stewart (ed.) (2005), *The Oxford Handbook of the Philosophy of Mathematics and Logic*, Oxford: Oxford University Press.

Tarski, A. (1983) *Logic, Semantics, Metamathematics*. 2nd ed. Ed. John Corcoran, Indianapolis: Hackett.

Tennant, N. (1996), “The Law of *Excluded Middle* is Synthetic *A Priori*, if Valid”, *Philosophical Topics* 24, pp. 205-229.

Troelstra, A. & D. van Dalen (1988) *Constructivism in Mathematics: An Introduction* Vols. I & II, Amsterdam: North-Holland Publishing.

Varzi, Achillle (2002), “On Logical Relativity”, *Philosophical Issues* 10, pp. 197-219.

Woods, John & Bryson Brown (2001), *Logical Consequence: Rival Approaches, Proceedings of the 1999 Conference of the Society of Exact Philosophy*, Oxford: Hermes Science Publishing.

University of Minnesota
cookx432@umn.edu

철학적 분석 31 (2014) pp.203-213

A Critical Note on Claude Panaccio's *Ockham on Concepts* (Aldershot: Ashgate, 2004. xi + 197 pp.)

Philip Choi

In his book *Ockham on Concepts*, Claude Panaccio suggests a strong externalist interpretation (SE) of Ockham's view on perceptual content. I argue that his SE is in conflict with his another interpretation of Ockham's view on conceptual similitude.

Keywords William of Ockham, Claude Panaccio, Concept, Mental Language, Externalism, Perceptual Content

Is it possible to understand past philosophies with contemporary philosophical notions? Peter Strawson once said yes to this question without hesitation. He said “no philosopher understands his predecessors until he has re-thought their thought in his own contemporary terms.”[1] Claude Panaccio’s *Ockham on Concepts* (henceforth ‘OC’) is an excellent work written from Strawsonian point of view on history of philosophy. In this dense book, Panaccio reconstructs William of Ockham (c. 1287-1347)’s theory of concepts in contemporary terms, and shows many interesting similarities that Ockham’s theory has with recent ideas in analytic philosophy, such as Jerry Fodor’s Language-of-Thought hypothesis and the Externalist movement promoted by Tyler Burge and Hilary Putnam.[2] Along with Marilyn Adams’s monumental work on Ockham[3], OC is by far one of the best books we have on Ockham’s philosophy.

There are two big projects in OC: one positive and the other negative. The negative project (Chapter 4-6) is to dethrone the so-called ‘Standard

[1] Peter Strawson, *Individuals: An Essay in Descriptive Metaphysics*, (London: Routledge, 1990), 10-11.

[2] In his recent works, Panaccio expands his externalist interpretation to Ockham’s epistemology. See Claude Panaccio and David Piche, “Ockham’s Reliabilism and the Intuition of Non-Existents,” in *Rethinking the History of Skepticism: the Missing Medieval Background*, ed. H. Lagerlund (Leiden: Brill, 2010), 97-118. According to his interpretation, Ockham’s epistemology has striking similarities with Alvin Goldman’s process reliabilism. In addition, there has been a great deal of discussion concerning the similarity between late medieval epistemology in general and contemporary reliabilism. See, for example, Dominik Perler, “Does God Deceive Us? Skeptical Hypotheses in Late Medieval Epistemology,” in *Rethinking the History of Skepticism: the Missing Medieval Background*, ed. H. Lagerlund (Leiden: Brill, 2010), 171-92; Eleonore Stump, “Aquinas on the Foundations of Knowledge,” *Canadian Journal of Philosophy* 17 (1991): 125-58; Jack Zupko, “Buridan and Skepticism,” *Journal of the History of Philosophy* 31, 2 (1993): 191-221.

[3] Adams (1987).

Interpretation' of Ockham's theory of concepts, according to which Ockham's Mental Language is a kind of logically ideal, stripped-down language. On this picture, most of our intellectual representations are complex mental expressions that can be reduced to few primitive absolute concepts.[4] With his careful examination of Ockham's texts, Panaccio—quite plausibly, I think—shows that the Standard Interpretation is misleading. Ockham's Mental Language, on his view, has some connotative concepts as its primitives as well, and it is not that kind of logically ideal, radically reductive language.

I think Panaccio's negative project largely succeeded. So, in what follows, I will mainly discuss his positive project in OC. More specifically, I will focus on one critical point, concerning Panaccio's strong externalist interpretation (henceforth 'SE') of Ockham's view on the content of perception—'intuitive cognition (*notitia intuitiva*)' in Ockham's words (Chapter 1): his SE is in conflict with his interpretation of Ockham's view on conceptual similitude (*similitudo*) (Chapter 7).

An intuitive cognition, according to Ockham, is an immediate apprehension of an external particular. In this respect, it corresponds to what contemporary philosophers call 'perception.' And there are two kinds of intuitive cognition: sensory intuition and intellective intuition. Sensory intuition is an apprehension of an external particular by our senses, e.g., seeing and hearing, while intellective intuition is a conceptual grasping of an external particular by the intellect.

Panaccio's SE is a thesis about the content of *intellective* intuitive cognition.[5] According to SE, the content of an intellective intuition is

[4] For a defense of the Standard Interpretation, see Adams (1987) and Spade (1975).

[5] One might ask: "Then, what about the content of sensory intuitive cognition?" Since only the content of intellective intuition is conceptual, sensory cognition

wholly determined by an external, causal relation that an intuitive cognition has with its object. His claim is strongly supported by Ockham's interesting thought experiment below:

> Even if an angel intuitively sees [another angel's] cognition of a certain singular [thing], and—we may suppose that—that angel also intuitively sees this singular thing, nevertheless he would not see that [that] cognition is of this singular [thing]. This is because if there were two similar [things], equally approximate to the intellect, and [another angel intuitively] sees one of them, he would not know whether this cognition is of one singular thing more than of another singular thing, provided that they [i.e., two singular things] are maximally similar. (*Reportatio*. II. q. 16; *OTh* V, 378-9)

Suppose there are two indistinguishable twins, say, John and Zorn. Gabriel is intuiting John, and Michael tries to find out the object of Gabriel's intuition by looking into his mind, i.e., his internal states. Can Michael know that the object of Gabriel's intuition is John? Ockham answers no, since Gabriel's internal states could be a similitude not only of John, but also of Zorn. To know the particular object of an intuition, one must know what causes that intuition. This is one central claim of SE, which I call 'Causal Determination Thesis':

(1) Causal Determination Thesis (CDT): the content of an (intellective) intuitive cognition is wholly determined by the causal relation that an intuitive cognition has with its object.

would either (i) be non-representational or (ii) have pre- or non-conceptual content. For a defense of (ii), see for example Dominik Perler, "Seeing and Judging: Ockham and Wodeham on Sensory Cognition," in *Theories of Perception in Medieval and Early Modern Philosophy*, eds. S. Knuuttila and P. Karkkainen (Dortrecht: Springer, 2008), 151-69.

In addition to CDT, Panaccio holds that, an external particular, which is the cause of an intuitive cognition, just is the content of that intuition. To use Panaccio's vocabulary, an intuitive cognition, *qua* a mental singular term in one's Mental Language, is a "direct designator that … refers to its object without the help of any form of description (OC, 13-4)" which is determined by internal features of a perceiver. This is another central claim of SE, which I call the 'Direct Reference Thesis':

> (2) Direct Reference Thesis (DRT): an intuitive cognition, *qua* a singular term in one's Mentalese, directly refers to its object without any descriptive content.[6)]

So far so good. There seems to be no serious inconsistency in SE, and SE seems to have strong textual evidence. Let us now focus on Panaccio's another interpretation. In the Chapter 7 of OC, Panaccio holds that Ockham never abandons the belief that concepts are similitudes (*similitudo*) or assimilations (*assimilatio*) of thing(s) that they represent. The following text seems to support Panaccio's claim:

> Then, I say that an intellection [including an intellective intuitive cognition] is similitude of the object just like a species would if it was admitted, and no more a similitude of one object than of another [similar] one. Thus, likeness is not the precise reason why [the intellect] understands one thing rather than another…For although the intellects assimilates equally to all these individuals…nevertheless it can cognize one of them determinately and not the other one. (*Reportatio* II. qq. 12-3; *OTh* V, 287-8)

The notion of assimilation or similitude is closely related to internal

[6)] I am not alone who analyzes SE into CDT and DRT. See Brower-Toland (2007).

features of an intuitive cognition, as we have seen in the above thought experiment: Michael cannot know the proper object of Gabriel's intuition only by looking into his mind since internal states of Gabriel could be a similitude of two indistinguishable twins. That is why Ockham appeals to the external, causal relation to explain the intentionality of intuitive cognition. Does this mean that Ockham entirely renounces the notion of similitude and explain the intentionality of concepts solely by causality? Many proponents of SE say yes[7], but Panaccio says no. On his view, Ockham never abandons the belief that concepts are similitudes. Thus, Panaccio seems to adopt the following thesis:

(3) Similitude Thesis (ST): an intuitive cognition, *qua* a similitude, can represent many objects that are maximally similar.

This is an interesting point. For ST seems to be in conflict with CDT and DRT: if ST is true, then, contra CDT, at least some of intuitive content, i.e., the content that makes an intuition a similitude of maximally similar things, would not be determined by an external, causal relation. Also, if ST is true, then, contra DRT, an intuitive cognition would have some descriptive content that makes an intuition a similitude of maximally similar things. For example, if my intuition of John contains some descriptive content, e.g., John's paleness, being tall, and so on, my intuition *qua* a similitude would represent things that are maximally similar to John, i.e., pale and tall things.

To resolve this inconsistency problem, Panaccio slightly revises his

[7] See for example King (2007) and Normore (2003). King holds that "although he [Ockham] preserves the traditional terminology, declaring that 'the act of understanding is the likeness of the object', it's clear that this is an empty formula (King 2007, 98)."

interpretation in his later works, e.g., Panaccio (2010), (forthcoming). He saves CDT and ST by abandoning DRT, i.e., an intuitive cognition directly refers to its object *without any descriptive content*. He grants that "Ockham does hold that an intuitive cognition internally incorporates some description of its object (Panaccio 2010, 244)." But this does not mean that he abandons DRT as a whole. He claims that "the likeness component⋯is not part of the *semantic* content of an intuitive cognition; rather it is *pragmatically* required to secure the correct reapplication of the concept (Panaccio 2010, 244)." So, for example, even if I represent John as pale when intuiting him and hence my intuition can represent all pale men, John's paleness is not part of the semantic content of my intuition; it does not determine what my intuition refers to.

Although Panaccio's modified view seems to resolve the tension between SE and ST, I think this view also has some problems. First, as Brower-Toland (2007) has shown, Panaccio's view is too coarse-grained to individuate intuitive cognitions of the same object. As is said, according to SE, intuitive content co-varies with an external particular which is the cause of an intuitive cognition. That is, without difference in the object, there would be no difference in intuitive content. However, this sits uneasily with Ockham's view on the relation between an intuition and perceptual judgment:

> Intuitive cognition is such that when certain things are cognized, one of which inheres in the other, or is distant from the other, or stands in some relation to the other, it is at once known by virtue of this non-propositional cognition of those things whether a things inheres or does not inhere, whether a thing is distant or not distant, and so on. (*Ordinatio* Prol. q. 1, a. 1; *OTh* I, 31)
>
> As he [the cognizer] approaches this visible object (say, a white thing), his

> vision of it is intensified and becomes clearer. And, accordingly, diverse judgments can be caused－for example, that the thing seen is a being, or a body, or a color, or a paleness, etc. (*Quodlibeta* I. 13; *OTh* IX, 76)

Here, Ockham claims that different intuitions of the same thing cause different perceptual judgments. But why would he think this? A natural answer would be that different descriptive content in intuitive cognitions of the same thing, e.g., John as being, John as body, John as colored, John as pale, makes the difference in perceptual judgments that immediately follow those intuitions, e.g., “John exists”, “John is a body”, “John is colored”, “John is pale”.[8]

Secondly, positing descriptive content explains perceptual indistinguishability much better than Panaccio’s interpretation. Consider the following example, which I borrow from Susanna Schellenberg:

> Imagine that Anna sees cup$_1$ at time t$_1$. Then she closes her eyes briefly and without her noticing cup$_1$ is replaced with the qualitatively indistinguishable cup$_2$. So when she reopens her eyes, she is causally related to a numerically distinct cup. Even though she cannot tell, her experiences before and after the cup was exchanged are of different objects. If she perceives cup$_1$ at time t$_1$ and cup$_2$ at time t$_2$, then her claim that the cup she sees at t$_2$ is the same as the cup she saw at t$_1$ does not have the status of knowledge, since the claim is false. (Susanna Schellenberg, “Perceptual Content Defended,” *Nous* 45, 4 (2011), 735)[9]

[8] This reasoning depends on what Jeff Speaks recently calls the Perception/Availability Principle, according to which “If two experiences differ in which thoughts they make available to the subject of the perception, then they differ in content (Jeff Speaks, “Transparency, Intentionalism, and the Nature of Perceptual Content,” *Philosophy and Phenomenological Research* 79, 3 (2009), 560).”

[9] Interestingly, Schellenberg’s example has a long history. Gyula Klima uses the very similar example in Klima (2009) when he discusses late medieval debates

In the example, Anna cannot distinguish two cups by her perceptual abilities, and hence she cannot notice the fact that one cup was replaced with another cup. If we posit descriptive content, this phenomenon can easily be explained by the sameness of the descriptive content between Anna's two intuitions: since Anna represents two cups as the same, she cannot distinguish them. Now, how can Panaccio's SE explain this phenomenon? It would be a hard problem since, according to SE, there is no common content between Anna's two intuitions; they would have different cups as their sole semantic content, no matter whether Anna distinguishes two cups or not.

To sum up, despite its initial plausibility, I think Panaccio's SE ultimately fails. But such criticism only slightly diminishes Panaccio's overall achievement that he shows throughout OC.OC is outstanding work, and recommended reading not only for anyone interested in Ockham's philosophy and late medieval theories of cognition, but also for anyone interested in contemporary philosophy of perception and cognitive science.[10]

on singular concept and singular cognition: "when I see horse, I have a singular cognition of this particular horse, which there I can name by a proper name expressing my singular concept of this particular horse? … Well, here is the problem … if in a blink of my eye someone (say God, to go directly to the top) swapped it for another, exactly similar one, I would not notice the difference. (Klima 2009, 69)" Klima holds that Nicole Oresme already suggested this kind of example－the case of two indistinguishable eggs－in the 14th century.

[10] Thanks to Jaekyung Lee for helpful comments.

References

Adams, Marilyn Mccord. (1987) *William Ockham*, 2 vols. University of Notre Dame Press.

Brower-Toland, Susan. (2007) “Intuition, Externalism, and Direct Reference in Ockham.” *History of Philosophy Quarterly* 24(3), pp. 317-335.

King, Peter. (2007) “Rethinking Representation in the Middle Ages: A Vade-Mecum to Mediaeval Theories of Mental Representation.” in H. Lagerlund (ed.), *Representation and Objects of Thought in Medieval Philosophy*, Ashgate. pp. 81-100.

Klima, Gyula. (2009) *John Buridan*. Oxford University Press.

Normore, Calvin. (2003) “Burge, Descartes, and Us.” in M. Hahn and B. Ramberg (eds.), *Reflection and Replies: Essays on the Philosophy of Tyler Burge*. The MIT Press. pp. 1-14.

Panaccio, Claude. (1990) “Connotative Terms in Ockham’s Mental Language.” *Cahiers d’épistémologie* n. 9016, pp. 1-22.

________________. (2010) “Intuition and Causality: Ockham’s Externalism Revisited.” *Quaestio* 10, pp. 241-252.

________________. (forthcoming). “Ockham’s Externalism.” in G. Klima (ed.), *Intentionality, Cognition, and Mental Representation in Medieval Philosophy*. Fordham University Press.

Spade, Paul Vincent. (1975) “Ockham’s Distinctions between Absolute and Connotative Terms.” *Vivarium* 13, pp. 55-76.

William of Ockham. (1967-86) *Opera Theologica (OTh)*, 10 vols. The Franciscan Institute.

__________________. (1974-88) *Opera Philosophica (OPh)*, 7 vols. The Franciscan Institute.

Department of Philosophy, University of Colorado, Boulder
philip.choi@colorado.edu

철학적 분석 31 (2014) pp.215-229

Whither Empiricism? An Essay Review of 'Scientific Representation' by Bas van Fraassen

Jeongmin Lee

This article is a belated review of the latest book by Bas van Fraassen. I discuss some main issues arising in the book and how they relate to one another. I also note how he redefines the empiricist tradition he inherits and how he subtlely transforms his previous empiricist position. I conclude with the provisional remark about whether his new position unduly imports some Kantian notions in that process.

subject philosophy of science, philosophy of art

Keywords representation, measurement, model, structuralism, empiricism

1. Introduction

Kant famously stated that it was the remembrance of Hume which awakened him out of his “dogmatic slumber.” I should confess that it was the reflection on the varied writings of Van Fraassen which had a similar impact on me. Not that my education was infused with some dogmatic metaphysics as Kant was with Lebniz-Wolffian school philosophy of his day. Nor do I have any critical resources of my own, which will allow me to evaluate impartially the ongoing debate between Van Fraassen’s empiricism and his opponents’s realism. Rather, as a student of physics who dabbled in the history and philosophy of science, it was my experience of reading him that awakened me to my *ignorance* about the nature of science. Science, according to Van Fraassen, was nothing like what I thought it’s like, and I’d like to emphasize this point for potential readers of this book, especially for philosophers who are keen on metaphysics. *Science is not what you think it’s like*. Unless you shed your a priori, metaphysical prejudices, you will never understand science. As far as I can see, that is the chief message Van Fraassen has tried to get across through his work, and *Scientific Representation* is the most systematic expression of his view available to date.

The book is based on his Locke Lectures of 2001. True to its namesake, Van Fraassen continues to explore and spell out a viable empiricist position, which has engaged him over thirty years. The publication of *The Scientific Image* in 1980 was a watershed event in philosophy of science, overturning the then prevalent realist consensus and at once offering an attactive empiricist alternative. Self-consciously carrying on the tradition of logical positivism, he has been a vocal critic of analytic metaphysics as well as scientific realism in such books as

Laws and Symmetry (1989) and *The Empirical Stance* (2002). While these books, though not lacking some constructive accounts, are largely polemical in their tone, his most recent book is the true successor to his 1980 book. With the main question, "what is scientific representation," firmly in his grip, he touches on as many related topics as representation in art, measurement and models, structure and realism, and not the least, the age-long philosophical distinction between appearance and reality. All in all, this book offers not just an up-to-date empiricist philosophy of science, but also a respectable continuation of empiricist philosophy since Locke.

2. Representation

Not to mislead potential readers, however, notice that 'representation' for Van Fraassen has little to do with the traditional discussion of mental representation. He goes so far as to claim that "mental representation is an oxymoron" (345, n.1). Representation in science as well as in the art is always by means of human-made artifacts, be it artistic creations or mathematical models. His account is about how these artistic and scientific artifacts represent intended targets in their respective domains, not about how our mind represents the external world. The reason he distances himself from any mentalistic notions is not hard to guess. As soon as you start talking about representation in mentalistic terms, you bring in a whole cluster of philosophical pseudo-problems like the mind-body problem, the problem of the external world, etc. that plagued early modern philosophy and that still plagues philosophy in the guise of the manifest image vs. the scientific image. As soon as the representation in question gets located in individual mind, *"[t]he problem initially faced* *in the sciences was thus transposed into one pertaining to mind and matter.”* (275, original emphasis) Rather, we should reformulate the whole problem of appearance vs. reality as the early modern philosopher-scientists like Copernicus and Galileo did. It was the problem of how scientific models (for them, heliocentric ones) represent the phenomena (in this case, the heavens) for human (earthbound) observers. It was this real problem of scientists that was subsequently replaced by the unsolvable pseudo-problems of modern philosophy.

A similar reservation about traditional philosophy also applies to his general account of representation. He lists and glosses in detail several key requirements any successful representation should meet. However, the account is not a philosophical “theory of representation.” (23) Readers familiar with his previous work may smell that the view of empiricism as “stance” (Van Fraassen 2002) is at play here. Anyhow, he builds upon Nelson Goodman’s earlier distinction between representation-of and representation-as and specifies a polyadic representation relation: A represents B as C for D. The key point here is that successful representation, artistic and scientific alike, trades on selective likeness as well as selective distortion. The respect in which such selection is made crucially depends on the use for which the representation in question is intended. Thus the representer’s intention is inseparable from the select contents of representation. This is another reason why mental representations do not count because they cannot be intentionally put to any possible use. Moreover, the users of the representation should be able to locate themselves with respect to it. To illustrate these points, take a metropolitan subway map. Here, the only likeness the map bears to the lay of the land is the topological connection between adjacent stations. The other land features such as distance, direction, elevation, etc. between stations are either significantly distorted or completely abstracted. Still, the purpose of the map is well served as long as you can count the number of stations left, the next transfer station, etc. More importantly, in order to use the map in those ways, you should be able to locate your place in it: “I am now here on the map.” Such *indexical* judgment or the need for self-ascription, while not being a part of the representation itself, is crucial for its application and won’t simply go away on pain of infinite regress(78-82 for arguments).

A closely related, but not identical, point holds for the perspectival character of representation. It is all the more evident in the technique of painting where the artist places his eyes, from which selective extraction and distortion is supposed to be made. We have such an artwork as Velasquez’s *Las Meninas*, which plays on subtle shifts of perspectives in and out of the painting and thus on the very perspective-bound character of visual representation. It is more or less the same situation in scientific representation. While mathematical models are not perspectival representations in themselves, in order for them to describe empirical phenomena, they must be related to the measurement outcomes, which are explicitly perspectival, as we will see.

All in all, in Part I of the book alone, Van Fraassen discusses judiciously a rich array of related subjects: likeness, distortion, intentionality, indexicality, perspectivity, etc. should all come into play in any nuanced account of representation. While his account is framed within his overall empiricist philosophy of science, this part can also be read as an independent contribution to the vast literature on representation.

3. Measurement and Models

The next part about measurement properly belongs to philosophy of science, but it also connects to his general account of representation in part I. Some readers may also recall the “measurement problem” familiar from quantum mechanics in this context, and this part certainly provides some necessary backgrounds for the dissolution of the problem in part IV. However, measurement is ubiquitous in all areas of representation. In the history of Renaissance art, the technique of perspectival drawing was chiefly an art of measuring. In science, measurement is all the more important because it gives scientists their “main initial access to the phenomena” (87). For Van Fraassen, scientific measurement is a form of representation, by means of which the target is located in a “theoretically constructed logical space” (2, 164-6). The act of measurement does not reveal any predetermined values. Instead of showing what the object is like in itself, measurement outcomes reflect what the object looks like under a particular measurement set-up. Hence both intentional and perspectival characters of representation figure indispensably in the act of measurement.

To continue his anti-empiricist theme, Van Fraassen regards modern-day scientific instruments as “engines of creation” rather than “windows on the invisible world.” In his memorable phrase, *“experimentation is the continuation of theory construction by other means”* (112, original emphasis). By generating phenomena that would have not been observed otherwise, scientific instruments have vastly expanded the scope of the observable phenomena, which is to be saved by highly mathematical models. Against misleading accusation that such phenomena are merely subjective experiences, Van Fraassen classifies them, together with naturally occurring phenomena like a rainbow, as *public* hallucinations in contrast to private ones like dreams. The former are available for public display and intersubjective confirmation, and their structure is to be saved by our scientific theories as much as the structure of 'real' objects.

As an avowed empiricist, Van Fraassen then owes us an account of how our scientific representation latches onto significant regularities in the phenomena. This he does meticulously in several steps, largely by creating some intermediate models and then embedding them within abstract theoretical models. Data models are summaries of measurement outcomes, and they are constructed from the raw data by our recording relative frequencies of salient outcomes. Surface models, on the other hand, are further "smoothing" of the data models, with relative frequencies being replaced by continuous probability measures. You may doubt whether such a distinction is universally applicable to all scientific data processing, but admittedly significant mathematical maneuver must be involved to make raw data amenable to theoretical analysis. Thus both data and surface models are already mathematically processed and thus have their own rudimentary structures. Now the question of empirical adequacy reduces to whether these structures are embeddable within theoretical models. The empiricist goal, to save the phenomena, is in many cases conveniently achievable by saving the appearances (measurement outcomes) or the data models. It is at this point that the realist worry gets acute because in most cases theoretical models, far from latching onto the real structures of the world, do not directly confront observable phenomena, but just already processed representations. Before dealing with the realist worry, however, there is already a formidable problem facing this view, to which we now turn.

4. Structure and Realism

By postulating empirical adequacy to be a match between two mathematical structures, Van Fraassen is committed to the structuralist view of scientific theories. The match in question is mathematical isomorphism between empirical substructures of a theory and the observable phenomena. If structures are all there are to be saved by theories, and they are describable only up to isomorphism, then a serious problem immediately besets such structuralism. It is well known that given a certain structure and a set, some relations can be defined on the set such that the set has that structure, the only constraint being on the cardinality of the set. Initially raised by Newman against Russell, the Newman problem is a radical form of an underdetermination argument against structural realism. However, the problem also has the following undesirable consequence for empiricists: any theoretical model having the cardinality of the target structure can be so organized as to be trivially empirically adequate. To get out of the dilemma, realists often postulate some privileged class of relations, with which we are directly acquainted (Russell) or which carve nature at its joints (David Lewis).

Van Fraassen's solution not surprisingly appeals to the role of indexical judgment in the application of representation. For Van Fraassen, there is no *pragmatic* difference between:

> a theoretical model M embeds the data model D, and
> a theoretical model M saves the phenomena represented by the data model D.

The reason for this tautology is illustrated with the following example. Suppose that a professor studies the growth of deer population in

Princeton. He constructs a theoretical model for it by reflecting various factors such as natural growth rate of the deer species, availability of food and habitat, human interventions, etc. The model fits very well the dotted graph (the data model) constructed from repeated measurements over some period of deer population in Princeton. However, how does we know that the model fits the actual growth of deer population, not just its graphic representation? For the professor who is a pragmatic empiricist, however, there is no difference between the two because it is *his* representation of the actual growth that matters. Someone else might plot a different graph and construct a different theoretical model, but essentially the same question will arise for that person. A kind of “hermeneutic circle” is broken and the “link to reality” is established when the person points to the graph and announces an indexical statement: *this* is the deer population in Princeton *for me*. Of course, you can dispute the practical procedure for data collection and graph construction, but that question is no longer the same as the original realist worry and does not presuppose any metaphysical claim about the reality of deer population.

Then how does this pragmatic tautology help to dissolve the Neuman problem? Once you accept the pragmatic tautology, the trivialization move does not go all the way through from one model to the next. It must come to an end when we *use* the representation by making indexical judgments to relate the model to the phenomena modelled. If you are still worried, try to think about it in the following way. In mature science, the contact with the real world is made only when the experimenter places the rods upon the target, so to speak. All the rest are representations upon representations. The narrow gap between readings and the target is all there is to the “loss of reality” objection. In this way, Van Fraassen’s radical demystification touches upon the venerable philosophical distinction between appearance and reality.

5. Appearance and Reality

Throughout his discussions, Van Fraassen makes the distinction between appearances and phenomena, the distinction that was absent in his previous work. Appearances are contents of measurement outcomes whereas phenomena are observable entities (objects, events, processes) on which such measurement is performed. Again, be careful not to associate such terms with individual sense perception. Phenomena such as the retrograde motion of Mars are publicly accessible and intersubjectively confirmable. Appearances are not different in that respect because though perspective-bound, they are obtained in a laboratory under a particular measurement set-up. What is different is that to save the phenomena does not entirely reduces to saving the appearances (though they are very close) because the former requires us to show how appearances derive from a particular perspective on the phenomena. For example, with his laws of planetary motion, Kepler saved the retrograde motion of Mars (phenomena) by predicting accurately how Mars appears to earthbound observers against the background of distant stars.

In my view, the distinction between phenomena and appearances is a clear advance over Van Fraassen's earlier term "observables." The observable/unobservable distinction has been controversial (rightly in my opinion) since its reinstallation by him, and the new distinction clarifies what is at stake much better than that all-encompassing term. With the help of the new distinction, the realist view of science, that mere empirical adequacy is not enough, is now reformulated as the *Appearance* *from Reality Criterion*. The requirement is that we must explain how the appearances derive from reality, or more exactly, that scientists must show the mechanism of how the appearances are produced as a proper part of the literally described reality. When the requirement is not met, then science is incomplete as Einstein's complaints about quantum mechanics illustrate. The notorious measurement problem of quantum mechanics is then understood as a realist demand for deriving definite measurement outcomes (appearances) from underlying quantum reality.

For Van Fraassen, however, such a demand is unreasonable for science. At least one scientific theory, now almost for a century, has faired well without respecting the above Criterion. The success of quantum mechanics is just one historical example, granted, but the central place of quantum mechanics in modern physics shows that Appearance from Reality Criterion cannot be an overriding aim of science. What the measurement problem shows is not that quantum mechanics is incomplete, but that the realist demand is unreasonable. Once you reject the Criterion, as empiricists recommend, the measurement problem is no longer a problem (in passing, I note this dissolution of the measurement problem is similar to Bohr's).

I personally thought that Bohr's insight has been largely lost in recent philosophical discussions of quantum mechanics, and Van Fraassen's recovery of the insight has much to recommend it. Though I doubt whether his solution has anything acceptable to realists, at least he succeed in shifting the realism vs. antirealim problematic from the observable/unobservable distinction to the Appearance from Reality Criterion. Whether the new problematic will invite a new round of controversies for many years to come, that I cannot tell, but realists have a formidable and ever stronger opponent to wrestle with, for sure.

6. Conclusion - Whither Empiricism?

Now setting aside quantum mechanics, philosophers may wonder where they can find any coherent epistemology in Van Fraassen's book. Certainly, we have multifarious accounts of representation in art and science, measurement and models, phenomena and reality, etc., but whither the central empiricist doctrine that experience is the one and only source of information? His anti-metaphysical "stance" as opposed to "doctrine" is certainly appreciated, though not agreed to, but is empiricism an epistemological position about the source of our knowledge including science? If Van Fraassen resists characterizing empiricism as a philosophical doctrine, then what is left to philosophy? Where does he stand on such issues dividing epistemologists as the internalism/externalism debate and the correspondence/coherence theories of truth?

As I understand, Van Fraassen's structural empiricism is a more sweeping and radical form of empiricism than hitherto envisaged by philosophers. His empiricism is a general view of what science is, not a view of what nature is like or how our knowledge is obtained or justified. It is an "empiricist view of empirical science" (Van Fraassen 2007, 367) but his empirical stance is a meta-scientific *attitude about* science, not any naturalistic philosophy on a par with empirical science. Accordingly, many traditional problems in metaphysics and epistemology are simply left behind. The empiricist doctrine of "experience" as the only "source" of knowledge contain too opaque terms for us to penetrate their meaning, especially when they are associated with individual psychic events. In Van Fraassen's enlightened empiricism, on the other hand, we have a radically demystified explication of how empirical adequacy is achieved in science, without any mention of experience, source, justification, and other problematic philosophical jargon. Certainly, this transformation of traditional philosophy began with logical positivists: Carnap conceived the logic of science as replacing the entire discipline of philosophy, and Reichenbach tried to transpose traditional metaphysical and epistemological problems into questions about scientific methodology. After eighty years, due partly to their failure and partly to the reactionary forces in analytic philosophy, their ambitious dream is yet to be realized, but Van Fraassen continues to explore a viable form of empiricism.

My final note should be cautionary rather than critical. What made me curious initially was Van Fraassen’s assimilation of the Sellarsian notion of “space of reasons” (84). Of course, the normative and justificatory concern of Sellars is not Van Fraassen’s, but his further view that measurement locates the target in a theoretically constructed logical space is exactly Sellarsian. As a Korean co-translator of Michael Friedman’s *Dynamics of Reason*, I could not fail to notice a similar assimilation is at work in Friedman’s Neo-Kantian view of science. If it is true that instruments as engines of creation generate hitherto unavailable phenomena and that “we and our efforts are part of nature” in that sense (96) what is there to stop Van Fraassen from becoming a transcendentalist instead of an empiricist? Given his view of empiricism as a stance, such labels may not amount to much unless they make practical differences about respective views of science, but I think it is indeed the case with transcendentalism vs. empiricism.

In the preface, Van Fraassen acknowledges a point of contact with transcentalists, and one author (Belot 2010) already queried the extent to which Van Fraassen’s view is Kantian. Although Van Fraassen (2009, 459) complains “transcendentalism seems to grant too much to the knowing subject," I think the role he assigns to indexical judgments is exactly parallel to the role played by Kantian a priori principles. On both accounts, indexical propositions or a priori principles are outside the purview of properly scientific models or empirical laws. According to the recent Neo-Kantian account of Reichenbach (Friedman 2001), a priori principles are not as absolutely fixed as Kant thought, but instead relativized to scientific theories. Relativized a priori principles therefore have nothing to do with in-built cognitive faculties of Kant, and rather should be conceived anew in each theory of mathematical physics as conditions of empirical possibilities. As much as I appreciate that Van Fraassen revives the historical problem of coordination, I regret he does not find attractive Reichenbach's solution to it (as Neo-Kantians do).

To be sure, transcendentalists and empiricists are united against dogmatic metaphysics of our time, and we further agree that the nature of science is explicable without recourse to metaphysical or epistemological preconceptions. Still the extent to which they agree or disagree on several key issues discussed in this review needs to be clarified. It is another virtue of Van Fraassen's latest book that it invites further work toward a rapprochement between the two recent versions of scientific philosophy.

References

All page references without author/year is to Van Fraassen (2008).

Belot, G. (2010) “Transcendental Idealism among the Jersey Metaphysicians”, *Philosophical Studies* 150, pp. 429-438.

Friedman, M. (2001) *Dynamics of Reason*, Stanford: CSLI Publications.

Van Fraassen, B. (1980) *The Scientific Image*, Oxford: Oxford University Press.

_____________. (1989) *Laws and Symmetry*, New York: Oxford University Press.

_____________. (2002) *The Empirical Stance*, New Haven: Yale University Press.

_____________. (2007) “From a View of Science to a New Empiricism”, in B. Monton (ed.), *Images of Empiricism: Essays on Science and Stances, with a Reply from Bas C. van Fraassen* (Oxford: Oxford University Press), pp. 337-383.

_____________. (2008) *Scientific Representation: Paradoxes of Perspective*, Oxford: Oxford University Press.

_____________. (2009) “Can Empiricism Leave Realism Behind? Toward a Dialogue with Transcendentalists”, in M. Bitbol, P. Kerszberg, & J. Petitot (eds.), *Constituting Objectivity: Trascendental Perspectives on Modern Physics* (Berlin: Springer-Verlag), pp. 459-479.

University of Seoul
philist@hotmail.com

윤리규정

제 1 조 (목적)

이 연구윤리규정은 한국분석철학회에서 발간하는 학술지인『철학적 분석』에 논문, 서평, 학술에세이 등을 투고하는 모든 학술 연구자 및 전문가(이하 ‘연구자’로 약칭)가 준수해야 하는 전반적인 사항을 정함을 목적으로 한다.

제 2 조 (연구윤리의 준수확인)

1. 『철학적 분석』에 게재하는 논문은 제3조의 연구윤리 규정을 지켜 작성하여야 한다.
2. 원고투고신청서 작성 시 윤리규정 준수 서약서에 동의하여 이를 원고 제출 시 반드시 함께 제출하여야 한다.

제 3 조 (연구자의 연구윤리)

1. 『철학적 분석』에 게재하는 논문은 다른 학술지 또는 간행물에 발표한 사실이 없는 독창적인 것이어야 한다.
2. 본 규정에서 ‘연구부정행위’는 다음을 말한다.
   (1) 타인의 아이디어, 연구내용・결과 등을 정당한 승인 또는 인용 없이 도용하는 표절 행위
   (2) 존재하지 않는 데이터 또는 연구결과 등을 허위로 만들어 내는 위조 행위
   (3) 연구내용 또는 결과에 대하여 학술적 공헌 또는 기여를 한 사람에게 정당한 이유 없이 논문저자 자격을 부여하지 않거나, 학술적 공헌 또는 기여를 하지 않은 자에게 감사의 표시 또는 예우 등을 이유로 논문저자 자격을 부여하는 행위
   (4) 본인 또는 타인의 부정행위 혐의에 대한 조사를 고의로 방해하거나 제보자에게 위해를 가하는 행위

(5) 기타 학계에서 통상적으로 용인되는 범위를 심각하게 벗어난 부정행위

3. 연구자는 자신의 논문이 게재된 이후 2항에 해당하는 연구부정행위가 발견되었을 경우 정정, 취소, 정오표 등 적절한 수단을 사용하여 오류를 바로잡는 조치를 신속히 취해야 한다.

제 4 조 (연구윤리 심의의결 기구)

1. 이 규정에서 정한 내용의 심의·의결은 본 『철학적 분석』의 편집위원회에서 담당하며, 그 위원장은 편집위원장이 겸임한다.
2. 편집위원회는 연구윤리의 위반과 관련하여 신고 되거나 자체적으로 인지한 내용에 대하여 규정에 의거하여 위반내용을 독립적인 지위에서 심의·의결한다.
3. 연구윤리의 위반과 관련된 회의는 편집위원 또는 편집위원장의 요청에 의해 이루어진다. 심의요청이 접수되면 편집위원장은 즉시 편집위원회를 소집해야 한다.
4. 위원회는 위원 과반수의 출석과 출석위원 3분의 2 이상의 찬성으로 의결한다.
5. 연구부정행위로 제보된 저자에게는 제보 내용을 통보하고, 소명자료를 제출하게 한다.
6. 편집위원회에서 필요하다고 인정될 때에는 관계자를 출석하게 하여 의견을 청취할 수 있다.
7. 편집위원회는 회의내용을 회의록으로 작성하여 보관하고, 심사 결과를 한국분석철학회 운영위원회에 보고해야 한다. 보고서에는 심사의 위촉내용, 심사의 대상이 된 연구부정행위, 심사위원의 명단과 심사철차, 심사 결정의 근거 및 관련 증거, 심사 대상자의 소명과 의견 청취 결과 및 처리 절차가 포함되어야 한다.

제 5 조 (연구부정행위의 제보 및 접수)

1. 제보자는 구술·서면·전화·전자우편 등의 가능한 모든 방법으로 제

보할 수 있으며 실명으로 제보함을 원칙으로 한다. 단, 익명으로 제보하고자 할 경우 서면으로 구체적인 연구부정행위의 내용과 증거를 제출하여야 한다.

제 6 조 (제보자와 피조사자의 권익보호 및 비밀 엄수)

1. 어떠한 경우에도 제보자의 신원을 직・간접적으로 노출시켜서는 아니 되며, 제보자의 성명은 보호 차원에서 조사결과 보고서에 포함하지 아니 한다. 단, 반드시 불가피한 경우 제보자의 동의하에서는 예외일 수 있다.
2. 제보자가 연구부정행위 제보를 이유로 징계 등 신분상 불이익, 근무조건상의 차별, 부당한 압력 또는 위해 등을 받은 경우 피해를 원상회복하기 위해 필요한 조처 마련에 적극적으로 나서야 하며 기타 제보자가 필요로 하는 조치를 취하여야 한다.
3. 연구부정행위 여부에 대한 검증이 완료될 때까지 피조사자의 명예나 권리가 침해되지 않도록 주의하여야 하며, 무혐의로 판명된 피조사자의 명예회복을 위해 최대한 노력하여야 한다.
4. 제보・조사・심의・의결 및 건의조치 등 조사와 관련된 일체의 사항은 비밀로 하며, 조사에 직・간접적으로 참여한 자들은 조사 및 직무수행 과정에서 취득한 모든 정보에 대하여 누설하여서는 아니 된다.

제 7 조 (연구윤리 위반에 대한 조치)

연구자가 연구윤리를 위반한 경우 편집위원장은 편집위원회에서 보고받은 내용을 기초로 한국분석철학회 운영위원회의의 결의를 통해 아래와 같은 조치를 결정하며 조치 내용은 중복될 수 있다.

1. 해당 논문을 학술지의 게재 목록에서 삭제하고, 해당 논문이 게재 논문임을 취소한다.
2. 한국분석철학회 홈페이지에 연구윤리위반 사실을 공지한다.
3. 한국학술진흥재단에 연구위반 사실을 통보한다.
4. 해당 연구자에게 향후 5년간 논문투고를 금지한다.

제 8 조 (부칙)

1. 규정에 명시되지 않은 사항은 편집위원회의 심의와 결정에 따른다.
2. 규정의 개정 또는 폐지는『철학적 분석』회칙 개정 절차에 준하여 시행된다.
3. 이 규정은 2008년 6월 15일부터 시행한다.

한국분석철학회 연구윤리 강령

1. 연구결과를 공표하거나 전문지식을 사회에 환원할 때 학문적 양심을 견지하고 지성인으로서의 책임과 의무를 다한다.

2. 연구 및 저술활동에서 저작권 침해, 표절, 부적절한 인용, 자료의 조작 등과 같은 비윤리적이거나 불법적인 행위를 하지 않는다.

3. 연구에 참여하는 연구원, 대학원생 및 연구 보조원의 권리나 인격을 침해하는 일이 없도록 하며, 이들이 기여한 정도에 따라 정당한 대우를 한다.

4. 연구자는 연구계약의 체결, 연구비의 수주 및 집행에서 윤리적 · 법적 책임과 의무를 이행한다.

5. 연구 활동에서 법률 및 본회 규정과 학계에서 권장하는 기본적인 연구윤리를 준수한다.

투 고 규 정

1. 『철학적 분석』은 자유 주제의 학술 논문, 비판이나 논쟁에 초점을 맞춘 토론, 그리고 '철학적 분석'에 부합하는 국내외 서적의 서평을 환영한다. 분량은 200자 원고지로 논문의 경우 100매 (20,000자 혹은 4,000단어) 내외, 토론과 서평은 각기 50매 (10,000자 혹은 2,000단어) 내외일 것을 권한다.

2. 원고는 일정한 마감 기한 없이 수시 접수하여 심사절차를 밟는다. 투고원고는 (한글 5.0이상의 버전 혹은 MS Word 등의) 워드프로세서로 작성하여 요약문 및 투고자 사항을 포함하는 파일을 『철학적 분석』편집실의 이메일 주소로 보내야 한다. 투고논문이 접수되면 편집실에서는 확인 메일을 보내고, 투고자는 접수사항을 확인하여야 한다.

3. 투고 논문의 필자가 2인 이상일 경우, 투고자들은 제1저자와 공동저자를 구분하여 명기하여야 한다.

4. 원고는 익명 심사에 적합한 방식으로 작성되어야 한다.
   (1) 제목(국문 및 영문)과 투고자의 성명(국문 및 영문), 소속(국문 및 영문), 우편 주소, 전화번호, 이메일 주소는 별지의 "투고자 사항"에 명기한다.
   (2) 투고논문에는 투고자의 신원이 드러날 수 있는 표현이나 내용이 포함되지 않아야 한다.

5. 철학적 분석에 게재되는 최종 원고는 다음의 양식을 따라 작성되어야 한다.
   (1) 논문 끝에 참고 문헌을 작성하되, 본문에서 인용 또는 참조된 문헌만을 수록한다. 동양어권의 논문은 " ", 저서 및 잡지는 『 』부호로, 서양어권의 논문은 " ", 저서 및 잡지는 이탤릭체로 표시한다.

예)

이명현 (2000) "'신문법' 그리고 철학", 『철학적 분석』 1호, pp. 1-14.

손병홍, 송하석, 심철호 (2002) "인공지능과 의식: 강한 인공지능의 존재론적 및 의미론적 문제", 『철학적 분석』 5호, pp. 1-33.

김효명 (1996) "흄의 필연성에 관한 인과적 고찰", 『인과와 인과이론』, 한국분석철학회편, 철학과 현실사, pp. 39-60.

정대현 (1994) 『필연성의 문맥적 이해』, 이화여자대학교 출판부.

Quine, W. V. (1951) "Two Dogmas of Empiricism", *Philosophical Review* 60, pp. 20-43.

Goodman, N. (1971) "Predicates without Properties" in French, P. et al. (eds.) *Contemporary Perspectives in the Philosophy of Language*, Minneapolis: University of Minnesota, pp. 347-348.

Hilbert, D. and P. Bernays (1968) *Grundlagen der Mathematik I*, Second Edition, Berlin: Springer-Verlag.

Kripke, S. (unpublished) "On Two Paradoxes of Knowledge", Transcript of recorded lecture given in Cambridge to the Moral Science Club.

(2) 주는 각주를 사용하되, 문헌은 필자와 발간연도 (필요할 경우에는, 쪽수) 등 최소한의 정보를 사용하여 언급한다.

(3) 논문 요약문 원고(국문 및 영문)가 첨부되어야 한다. 요약문 원고는 (i) 필자 자신에 의한 주제 분류, (ii) 필자가 선정한 5단어 정도의 주요어 또는 검색어, (iii) 500자 미만의 요약문으로 구성한다. 주제 분류는 통상의 형이상학, 인식론, 윤리학, 논리학, 사회철학, 언어철학, 심리철학, 과학철학, 미학 등 일반적인 방식을 따르며, 인물을 중심으로 하는 분류도 가능하지만 분류 주제의 수는 2-3개 정도로 한다. 주요어 또는 검색어는 논문에서 논의되는 주요 개념들이나 주제들 또는 인물들을 언급하되, 다섯 개 내외로 한다.

(4) 논문의 저자의 소속과 연락처(e-mail)는 원고의 끝 부분에 기재한다.

편 집 규 정

1. 『철학적 분석』은 한국분석철학회의 회지로서 일 년에 두 번(6월 30일과 12월 31일) 정기적으로 발간된다.

2. 『철학적 분석』은 두 가지 목표를 지향한다. 첫째, ‘분석’이라는 방법론적 단어를 넓게 해석하여 선명한 철학적 문제 제기와 엄밀한 논의 전개로 이루어진 글은 모두 ‘철학적 분석’의 논문이라 할 것이다. 그리하여 이 학술지는 특정 지역이나 시대의 철학적 전통에 매이지 않고 이 땅에서 철학하는 모든 동료들과 더불어 생각하는 철학적 마당이다. 둘째, 한국철학 공동체가 그동안 세계 문맥과의 연대에 치중하다가 한국 문맥과의 관련에 소홀하지 않았나 생각하여 균형을 찾고자 한다. 이 균형은 한국 철학자들 간의 상호 토론을 통하여 얻어질 수 있을 것이다. 달리 말하여, 이 학술지는 분석철학 전공이 아니어도 어떤 한국학자의 철학에 대한 논의를 엄밀하고 비판적으로 쓴 논문을 환영한다.

3. 심사 절차:
   (1) 투고된 논문은 편집실에 투고 논문의 접수가 완료 되는대로 심사절차에 들어간다.
   (2) 편집위원회는 투고논문의 내용과 관련된 연구 업적이 있고 공정한 심사가 가능하다고 판단되는 3인의 심사위원을 선정하여 이들에게 논문의 심사를 의뢰한다. 심사위원은 투고논문의 창의성, 엄밀성과 명료성, 치밀성과 완결성, 문제의 중요성과 논문의 기여도, 국내 및 국외 선행연구의 반영도 등을 항목별로 평가하고 종합적으로 게재여부를 판정하여 “심사결과표”와 “심사의견서”를 작성한다.
   (3) 편집위원회에서는 3인의 심사위원이 작성한 심사결과표와 심사의견서에 토대하여 게재여부를 최종 결정한다. 3인의 심사위원 중 2인 이

상이 게재가능으로 평가하였을 경우 게재하는 것을 원칙으로 한다. “심사의견서”는 게재여부와 상관없이 투고자에게 보내진다.

4. 게재 절차 및 요건:

(1) “게재 가능” 판정을 통보받은 투고자는 심사의견서를 참조하여 투고규정 5항 등의 지침에 따라 수정 또는 보완된 최종 원고 파일을 편집실의 이메일 주소로 전송하여야 한다.

(2) 저자가 논문게재비 지출을 허용하는 연구비를 지원받고 연구비 지원사항을 논문에 표기할 경우, 게재논문에 다음 (i)과 (ii) 중 해당되는 금액이 논문게재비로 부과된다.

(i) 수혜연구비에 논문게재지원비가 명시된 경우: 명시된 금액.

(ii) 수혜연구비에 논문게재지원비가 명시되어 있지 않은 경우: 30만원.

(3) 게재 논문의 별쇄본을 필요로 하는 경우, 필자는 최종 원고를 제출할 때 필요한 별쇄본 부수를 요청하여야 하며, 별쇄본 제작비용은 필자가 부담하는 것을 원칙으로 한다.

(게재비 납부 계좌: 신한은행 100-026-989940 예금주 김세화(한국분석철학회).)

(4) 투고자의 논문은 게재와 동시에 논문의 저작권이 한국분석철학회로 이양된다.

Submission Guidelines

1. *The Philosophical Analysis* is periodically published by Korean Society for Analytic Philosophy twice a year on June 30 and December 31.

2. We welcome articles and book reviews in Korean or English by philosophers of any country on any aspect of philosophy relevant to ‘philosophical analysis’

3. We accept submissions throughout the year. All submissions should be made by sending to philosophical.analysis.editor@gmail.com an email attached with the manuscript file in the format of HWP or MS Word or PDF. All subsequent correspondences such as submission confirmation will be sent via email.

4. In case the manuscript is written by several authors, the first author should be designated as such.

5. The manuscript must be prepared for double-blind refereeing.
   (1) The title of the manuscript and the personal information including your name(s), affiliation(s), postal code(s), telephone number(s), and email address(es) should be provided on the submission message or on a separate cover letter.
   (2) The manuscript must not include any contents or expressions that would reveal your identity.

6. Guidelines for reference list, note, and abstract

(1) The reference list must be put at the end of the manuscript; and all and only those bibliographical materials quoted and referred to in the manuscript must be included. Book and journal titles must be italicized and double quotation marks must be used for titles of journal articles.

e.g.)

Quine, W. V. (1951) "Two Dogmas of Empiricism", *Philosophical Review* 60, pp. 20-43.

Goodman, N. (1971) "Predicates without Properties" in French, P. et al. (eds.) *Contemporary Perspectives in the Philosophy of Language*, Minneapolis: University of Minnesota, pp. 347-348.

Hilbert, D. and Bernays. P. (1968) *Grundlagen der Mathematik I*, Second Edition, Berlin: Springer-Verlag.

Kripke, S. (unpublished) "On Two Paradoxes of Knowledge", Transcript of recorded lecture given in Cambridge to the Moral Science Club.

(2) For annotation, you must use footnotes. For an in-text reference, you only have to make reference to the author, the year of publication, and, if necessary, the page number(s).

(3) The submission must include 2-3 subject categories, (more or less) 5 key words, and an abstract. Though standard subject categories such as metaphysics, epistemology, ethics, logic, political philosophy, philosophy of language, philosophy of psychology, philosophy of science, aesthetics, etc. are recommendable, it is also possible to invent your own subject categories that you think most suitable to classify your discussion. Key words should mention main concepts, topics, or figures of the manuscript.

7. Refereeing process

(1) The manuscript will undergo a refereeing process soon after submission confirmation.

(2) The editorial board selects 3 referees for review. They evaluate the manuscript in terms of its originality, clarity, completeness, significance, contribution and so on. They will give a referee report with the opinion as to its publishability to the editorial board.

(3) The editorial board makes the final decision as to the publication of your manuscript based on the referee reports. The referee reports will be passed on to you irrespective of the editorial verdict.

8. Publication process

(1) When your manuscript is accepted for publication you must send its final draft to the editorial office.

(2) The copyright of the published manuscript will be transferred to Korean Society for Analytic Philosophy after publication.

9. If your research has been supported by a research grant that provides funding for the fee for publishing an article in an academic journal, and if you need formally acknowledge this support in your article, then a publication fee, which is the maximum amount allowed by your research grant, will be charged.

10. You may buy off-prints of your article at a reasonable price, in which case you should notify us of the number of off-prints you want when you submit the final draft.

철학적분석

발행처 : 한국분석철학회
발행인 : 정인교
편집인 : 박우석

인쇄처 : 한국학술정보(주)

2014년 12월 30일 초판 1쇄 인쇄
2014년 12월 31일 초판 1쇄 발행

題字 : 이명현
표지그림 : 엄정식