# INTRODUCTION TO PHILOSOPHY

***The Complete Notes***

**(Spring 2011)**

**MEETING 1**

## 1. The Field of Philosophy: Areas, Questions, Distinctions and Some Terminology

### 1.1. *Areas and Questions*

Perhaps the only simultaneously easy and accurate way of introducing philosophy is by listing some of the questions that philosophers try to answer, and saying a bit about how philosophers think (perhaps wrongly!) that they can satisfactorily answer them. So, I'll start by observing that the questions of philosophy fall into the following broad areas:

**Epistemology**. This branch (my area of specialization) is concerned with questions such as:

- What is knowledge? (The Greek word for knowledge is *episteme*; whence *epistem-ology*.)
- How do we know what we know, and how much do we know?
- How is knowledge different from mere true belief? When someone knows the truth of some claim, does he have to have good reasons for believing that claim? Or does his belief in the claim merely have to be produced by a reliable cognitive process (e.g., visual perception), and not be defeated by countervailing evidence?
- What are good reasons for belief, anyway? Are all proper reasons for belief *evidential*? Does justified belief always require good reasons for belief?
- What makes for good or legitimate inquiry? Are, for instance, philosophers today engaging in a good form of inquiry, or is there something defective about the way philosophers try to answer the questions that interest them?
- What is the *structure* of knowledge and justified belief? Am I only justified in believing that there is a table here *because* I'm antecedently justified in believing that I'm having a kind of visual experience that happens to be *reliably correlated* with worldly matters? Do all my beliefs about worldly phenomena have to be based on absolutely certain, indubitable beliefs about my own mental states to count as *fully justified*?

**Metaphysics**. This branch (my favorite as an undergraduate) is concerned with this question:

- If we wanted to provide the simplest fully accurate and complete description of reality, what *sorts* of things and properties would we have to talk about in providing it?

This big and difficult question subsumes a number of other smaller but also quite difficult questions, which then generate some further branches of philosophy:

  - Would we have to talk about *mental* phenomena like *beliefs* and *experiences* in giving the simplest fully accurate and complete description of reality? Or do mental phenomena *reduce* to physical phenomena? For instance, is what it is to have a visual experience *nothing more* than just being in some brain state?
    - A sub-branch of metaphysics called the **philosophy of mind** is concerned with questions like this.

  - Would we have to talk about *agents* who can *perform actions of their own free will* in giving the simplest fully accurate and complete description of reality? Or are there no facts about free agency, and the very idea is just an illusion? Or, more interestingly, do facts about free agency *reduce* (without being *eliminated*) to simpler mental phenomena, which in turn reduce to purely physical phenomena?

Is the world as described just by fundamental natural sciences like physics compatible with the familiar social world in which we believe ourselves to be?

- A sub-branch of the philosophy of mind (and hence of metaphysics) called the **philosophy of action** is concerned with questions like this.

- Would we talk about supernatural beings like God in providing a fully accurate and complete description of reality? If so, how could we ever know that there are such beings? Can we prove that there are, or do we have to take it on faith?
  - A sub-branch of philosophy that partly intersects with metaphysics called the **philosophy of religion** is concerned with questions like this.

- Would we talk about familiar perceptually observable properties (called "secondary qualities" by Locke) like *color* and *texture* and *smell* if we wanted to provide the simplest fully accurate and complete description of reality? Or are there really no such properties, and we simply wrongly *project* them onto the world? More interestingly, if there really are such properties, do they just reduce to facts about how simpler physical objects are disposed to cause people to have certain experiences?

- Would we talk about *causal relations* and *laws of nature* in providing the simplest fully accurate and complete description of reality? Or do apparent facts about causal relations and laws of nature just reduce to *mere regularities* and *patterns of events* in the world? (Hume, whom we'll read, said 'yes' to the second question.)
  - A sub-branch of philosophy that partly intersects with metaphysics called the **philosophy of science** is concerned with questions like this.

Metaphysics is probably the biggest area of philosophy, and has a lot of other branches of philosophy as parts. As you can see from the questions listed above, much of metaphysics is concerned with questions of *reduction*. Since metaphysicians aim to offer the *sparsest* description of the fundamental ingredients of reality, they often want to see how they can explain one kind of phenomenon (e.g., the mental, the causal, colors and smells, etc.) entirely by reference to a more fundamental kind of phenomenon (e.g., the physical or the chemical).

Of course, one worry is whether it's possible to do this without entirely *getting rid* of the former phenomenon: can we, for instance, really understand the world as just physical and deterministic without simply *eliminating* such apparently familiar facts as free will? This is a question that strikes many outsiders to philosophy as almost certainly having a negative answer: we'll see about that! (But some metaphysicians *love* elimination, and wouldn't see this as an objection to their quest for reduction [*Insert funny anecdote about Kit Fine at Ted Sider's metaphysics seminar at NYU*].)

**Value Theory**. This last major branch of philosophy is concerned with questions of value.

- Some of the questions here are about *moral* value, and the branch of philosophy that deals with these questions is **ethics**. Ethics itself divides into three branches. A branch often called **normative ethics** asks mid-level theoretical questions such as:
  - What makes an act right or wrong? Is the rightness or wrongness of an act entirely determined by the nature of the *consequences* it has? Or are some acts right or wrong regardless of their consequences?
  - For those acts that *are* made right partly by their consequences, what features of the consequences of acts are fundamentally relevant to

rightness? Are an act's implications for the *well being* of everyone in the world the only thing that matters? What is well being, anyway?

- Another branch of ethics called **applied ethics** or **practical ethics** is less theoretical. It is concerned with very specific questions about right action, like:
  - When, if ever, is abortion morally permissible?
  - When, if ever, is preemptive war morally permissible?
  - When, if ever, is assisted suicide morally permissible?
- The last branch of ethics is the most theoretical, and is called **metaethics.** Unsurprisingly, metaethics heavily intersects heavily with metaphysics. This branch of ethics asks very high-level, abstract questions like the following:
  - Are moral claims objectively true? Or do they simply *express* the emotions or practical dispositions of the people who assert them?
  - If moral claims are objectively true, how is their truth explained by or related to purely descriptive truths about the world? In Hume's words, is an 'ought' claim ever derivable from an 'is' claim?
  - How is moral judgment related to *motivation*? Is it necessarily the case that, if Jones says that killing is wrong, Jones is motivated to some degree to avoid killing people? Or could someone sincerely judge that A-ing is wrong without being even slightly disposed to avoid A-ing? If someone could, would that person have to be *irrational*?

- There are other parts of value theory concerned with *non-moral* value. One of the other big parts is **aesthetics**, which is concerned with the value claims that we make about art, music, literature, and so on. Sadly, we won't be getting into any aesthetics in this class, but it's a very fun and interesting area, and if you want to take a class on it, there are great people in our department who often teach it.

There are some other stray areas of philosophy I haven't mentioned here, but these are all the major ones we'll be getting into in this class. (Most of the areas I haven't mentioned are "philosophy of X" areas: philosophy of language, philosophy of physics, etc.)

### 1.2. *Philosophical Method and Some Terminology*

Besides the distinctively high-level and abstract questions that philosophers care about, the other thing that is often said to distinguish philosophy as a field from most other fields is its *method*.

Surprisingly to outsiders, philosophers have until quite recently acted as if they can do their job with nothing more than an armchair and maybe some other friends with whom they can argue. There is a distinction couched in Latin terminology often used by philosophers worth introducing here: the distinction between *a priori* and *a posteriori* knowledge. *A priori* knowledge is knowledge that you can get without any recourse to experience, and indeed just by engaging in *pure reflection* on the nature of your concepts and your intuitions about imaginary cases. *A posteriori* knowledge is knowledge that you can get only by *experience*, and by engaging in some kind of interaction with the world. Traditionally, philosophy has been viewed by philosophers as an *a priori* discipline: the questions of philosophy are, if knowable at all, knowable *a priori*.

Not all people are happy about this approach to philosophy and the presuppositions it makes, but it's the approach we'll be seeing throughout most of this class. A few of the particular philosophers we'll be discussing (e.g., David Hume) are skeptics about *a priori* knowledge, and perhaps ironically so, since they argue for their skepticism by armchair methods! But all of the philosophers we'll be discussing rely in most cases on nothing but pure thought to answer their

questions. Whether this is a legitimate form of inquiry is an open question, and one that people in the Rutgers philosophy department somewhat famously question (e.g., self-proclaimed "experimental philosophers" like Steve Stich). For most of the class, we'll just be taking it on trust that it's legitimate, and see where we can go on this assumption. If you don't like this assumption, that doesn't mean you shouldn't take this class: it may very well mean that you should, so that you can *then* take a more advanced philosophy class that attacks the assumption!

If we treat philosophy as an *a priori* discipline, how are we going to go about answering its hopefully fascinating and certainly difficult questions? We'll be using two methods. One is *deductive reasoning*. In pretty much every class, we're going to unpack the thoughts of historical and contemporary philosophers in neat arguments that have formal structures like the following:

**Argument Form I**

1. If A is true, then B is true.
2. B is false.
3. So, A is false.
4. If A is false, then C is true.

---

5. So, C is true.

The first four claims in this schematic argument are called **premises**, and the last is called a **conclusion**. If the philosophers we read are doing their job extremely well, the particular instances of argument forms like this will have a couple of nice properties.

One property is **validity**. A valid argument is one such that, ***necessarily****, if its premises are true, its conclusion is true.* The argument form I've just sketched has this property: if we substitute in actual sentences for 'A' and 'B', (1 – 4) cannot all be true unless (5) is also true.

The adverb 'necessarily' is quite crucial. Some invalid arguments have true premises and a true conclusion. (Example: (i) If I have gloves, then I have hands. (ii) I have hands. (iii) So, I have gloves. This is **not** a valid argument, and is instead an instance of an invalid argument form called **affirming the consequent**. Nevertheless, its premises are true, and so is its conclusion, and it's also true as a matter of *contingent fact* that if its premises are true, so is its conclusion: it's just not ***necessarily*** *the case* that if its premises are true, so is its conclusion, and so it's invalid.[1])

In fact, Argument Form 1 splits into two parts, each of which contains a valid inference. The move from (1 – 2) to (3) is an instance of a valid argument form called ***modus tollens***, whereby we reason from a **conditional** claim ("If A, B") and the falsity of its **consequent** (in this case, B), to the falsity of its **antecedent** (in this case, A). The move from (3 – 4) to (5) is also an instance of a valid argument form called ***modus ponens***, whereby we reason from a conditional claim and the truth of its antecedent (in this case, "A is false") to the truth of its consequent. *Modus tollens* and *modus ponens* are some of the most basic argument forms we'll be seeing, and most complex arguments can be broken into steps that apply one or the other of the two forms.

---

[1] Terms like 'necessarily' and 'contingently' that apply to the *way* in which a claim is true or false are called **modal** terms. I'll be using the word 'modal' when we come to Aquinas later today, so bear this in mind.

Another nice property that arguments ought to have (but, alas, don't always have) is **soundness**. A sound argument is a valid argument with true premises. Here is a rather dull sound argument that is an instance of Argument Form 1 that we'll apply to a fish I'll call Fishy:

**Dull Instance of Argument Form 1**

I. If Fishy has hands, then Fishy has fingers.
II. Fishy doesn't have fingers.
III. So, Fishy doesn't have hands.
IV. If Fishy doesn't have hands, Fishy doesn't have palms.
---
V. So, Fishy doesn't have palms.

Why have I subjected you to such a dull argument? Because sound arguments that have non-dull conclusions are *really hard* to construct. This is an instance of the more general principle that doing good philosophy is really damn hard.

Arguments can have other virtues. Perhaps the most important virtue is having premises that are *intuitively plausible*. This brings up the other key component of philosophical methodology, at least when philosophy is viewed as an *a priori* discipline: appeals to intuition.

One cannot argue for *everything* in doing philosophy properly unless it's impossible to do philosophy properly and thereby achieve results. Perhaps it is actually impossible to do philosophy properly and thereby achieve results, but let's be optimistic for a moment. If we are ever to actually embrace the conclusion of some philosophical argument, we're going to have to be comfortable with embracing the premises. Sometimes we'll want to argue for the premises. But eventually we're going to have to stop somewhere, even if only because of our cognitive limitations. The nicest place to stop is on a premise that is *obvious* or *just plain attractive*.

This is in fact rare, but there are some seeming examples. Suppose an unmanned trolley is speeding down the tracks and, if uninterrupted, will run over and kill two people. The only way I can stop it is by pushing a large man off the bridge on which I'm standing, and I know that no one else will stop the trolley. I decide to push him off the bridge. It's highly intuitive to say: "That is clearly wrong". Many would go so far as to say that this is unquestionably obvious. And this isn't something that seems to require argument: it's permissible to simply assert it. Perhaps it's still false, but it's going to take a lot to convince ordinary people that it's false. And this actually may seem to achieve an important result: if we accept this premise, and we see that certain simple forms of consequentialism in ethics entail that it is false, we can know that these theories are false. If we do know that, part of the basis of our knowledge will be *intuition*.

Alas, the history of philosophy is rife with appeals to claims that just aren't intuitively obvious, or that are arguably false and rest on unreliable intuitions. We'll now turn to some illustrations as we review Aquinas's arguments for the existence of God.

## 2. Aquinas's Ways

Since there was quite a bit of participation in assessing Aquinas's arguments in the lectures, I want to try to summarize some of the key criticisms that people brought up so that people remember them. So, let's turn to a quick review of these arguments.

## 2.1. *The First Way*

As we've seen, Aquinas's first argument for God's existence goes like so:

1. Some things change.
2. Change is a transition from potentiality to actuality in some aspect.
3. Nothing can go from potentiality to actuality in some aspect except by means of something that actually exists.
4. The same thing cannot be both potentially X and actually X.
5. So, nothing can move itself, assuming movement is a kind of change.
6. So, everything that is moved is moved by something else.
7. This chain of movement that extends backwards cannot extend backwards infinitely.
8. So, there is a first existing step in the chain – i.e., a "first mover".
9. And this first mover is God.

This argument is problematic in many ways. One of the most obvious ways – and a way that is indeed shared with the rest of Aquinas's arguments – centers on the step from (8) to (9). God, at least as Aquinas is understanding him, is supposed to be a being with a mind (and a gender, too, which seems a little dubious, but let's go with tradition for the sake of argument). He's a being with other properties, too: benevolence, omniscience, and so on. It seems, however, like claim (8) could be true even if no entity with a mind and all these other properties existed. The first mover could, it seems, just be a strictly physical event at the dawn of time, like the Big Bang. As long as we're willing to grant that the Big Bang isn't God, it seems like the move from (8) to (9) is clearly invalid. Since we'll come across this problem in the other arguments, let's give it a name: I'll call it the **Additional Properties Fallacy**, since Aquinas is inferring that the first mover must have many additional properties that the argument just doesn't establish it to have.

The other ways in which the argument is flawed are a bit more arcane, partly because the premises themselves are arcane. (Some of the premises do seem intuitively or factually plausible: (1), (2), and (3) are all, I think, quite plausible.)

A part that bothers me is the bare assertion of (4) in the argument. This premise seems doubtful for reasons that some people brought out in the lecture. Some things that have a property (e.g., hotness) to some degree might have it to a greater degree: something that is hot could be hotter. But part of the explanation of what makes something have the property to a greater degree will be its potentially having that property to that degree. And something can't potentially have a property to *any* degree without potentially having that property *period*. How could something be potentially hot-to-degree-X without being potentially-hot? That seems incoherent. If so, it *just follows* that something that is actually hot that is also potentially hotter is also potentially hot.

Now, I grant that this conclusion can *sound* odd. But the reason it *sounds* odd is just that we don't normally *say* that something is potentially X unless we also believe that it is not actually X. But from the fact that we don't normally say one thing unless we believe that something else is false doesn't mean that the truth of the first *entails* the falsity of the second. Consider an example. Suppose I say: "There are ten people in this room." Normally, I wouldn't say this if I didn't also believe that there were *only* ten people in this room: after all, we usually expect speakers to be *maximally informative*, and I would not be telling you the whole truth of which I was aware if I said that there were ten people and yet believed that there were thirty people in the room. But, clearly, if there *were* thirty people in the room, there would also be ten. But if there were *only* ten, there obviously couldn't be thirty. So, sometimes I can say one thing – call it A – *only when* I also

believe some other thing – call it B – even though A doesn't entail B. Here, let A = that there are ten people in this room, and B = that there are not thirty people in this room. A doesn't entail B, though I normally *suggest* B when I say A.

So, in short, it seems like there is a good argument against (4), and that the only reason for shrinking from the rejection of (4) is that it happens to "sound bad" for reasons that are irrelevant to its actual truth or falsity (reasons having to do with norms governing communication, such as the norm that one should assert the most informative claim possible). Once we reject (4), there's no reason to move from (4) and the claims that precede it to (5). And this actually seems like a good thing, since (5) itself seems false: surely some things can move themselves, since it certainly looks like *people* can move themselves!

The last clear thing that's problematic about the argument is that no good argument seems readily available for (7), and (7) indeed seems to come from nowhere. Of course, it is hard to *imagine* what it would be like for there to be a series of changes occurring now that weren't initiated by anything in the past. But it's hard to imagine many facts that are demonstrably true. I still can't wrap my head around the fact that there are infinitely many distinct orders of infinity (so that aleph-null, the lowest order of infinity (and the size of the set of natural numbers), is actually *smaller* than aleph-one, which is the size of the set of real numbers), and around the fact that the number of even numbers is the same as the number of natural numbers (both being aleph-null, the first order of infinity). But these claims are true, and indeed mathematically provable and universally accepted as quite *trivial* results in set theory! The idea of an infinite regress of change might be similar: indeed, we have to accept that it is if certain models of the Universe (where a Big Bang was preceded by a "Big Crunch" and so on indefinitely) that are taken quite seriously in cosmology are even possible.

### 2.2. *The Second Way*

Many of the same problems that plagued Aquinas's first argument also undermine his second, which, as you'll recall, goes as follows:

1. There is a cause of everything.
2. Nothing can be the cause of itself, because causal sequences are temporally extended (so that the cause of A must temporally precede A).
3. There can't be an infinite regress of causes.
4. So, there must be a first cause.
5. That first cause is God, and so God exists.

The move from (4) to (5) is another case of what I called the *Additional Properties Fallacy* last week. As I said before, God, as a Catholic theologian like Aquinas wants to understand him, is supposed to have many properties that the premises in this argument don't establish him to have. God is supposed to be benevolent, omnipotent, omniscient, and so on. And it's compatible with (1-4) that the first cause is malevolent, not omnipotent, not omniscient, and so on. The first cause might just be the Big Bang, and this argument doesn't establish that it isn't. For this reason, (5) doesn't follow from (1-4), and the argument is invalid.

Claim (3) is problematic for the same reasons we saw last week that undermined premise (7) in Aquinas's first argument: there's just no good argument for it, and it's not at all self-evident or obvious. Of course, Aquinas does seem to argue for the no-infinite-regress claim in presenting the Second Way, but his argument seems clearly circular, at least as far as I can make it out. So,

why accept (3)? The only reason I can see is that it's *hard to imagine* an infinite regress of causes. But it's hard to imagine the truth of most claims involving the infinite, and there are even some *provable mathematical results* involving the infinite that are highly counterintuitive, and persistently so. So, I don't think we can just rely on a brute appeal to intuitive difficulty of imagination as an argument for (3). And so there seems to be no clear good reason to accept (3).

This argument is even more clearly problematic for a different reason. As someone adroitly noted in the lecture, the intermediate conclusion (i.e., step (4)) is inconsistent with the conjunction of the first premise and the second premise: if there is a first cause, it would seem to be uncaused, which contradicts the claim that there's a cause of everything – this, at any rate, would be so if something couldn't cause itself to exist, and that's ruled out by (2), since causes and effects must be temporally separated, and it seems obvious that before something is caused to exist it doesn't exist. So, the argument looks *necessarily unsound*!

What about the other premises? Well, as I already suggested, we'd have to reject (1) if we believed in a first cause and in the principle that nothing is its own cause. So, even setting aside God, we'd have to do this if we believed in the Big Bang as the proper beginning of the Universe and rejected self-causation. Perhaps we're supposed to accept that God is special in being the only self-causing thing. God's supposed to have lots of special properties, so why not this one? But this looks like an attempt to explain the obscure by appealing to the more obscure (since self-causation really looks incoherent if (2) is true).

There is a different and simpler reason for rejecting (1). These days it's becoming standard to believe in *fundamental physical probabilities*: for some physical events, it's just an irreducible, brute fact that these events occur with a probability less than 1. People have believed this about radioactive decay: there is a chance that some unstable atomic nuclei will eject some of their constituents (e.g., electrons), and this chance is unexplained by any more fundamental physical facts. I don't know if *precisely* this is still believed. But I know it is widely accepted that, at *some* subatomic level, there are irreducibly chancy events. Insofar there is no way to get rid of the *randomness* of these events by further explanation, they would seem in a deep sense to be *uncaused*. If that's true, we can cast doubt on (1) on empirical grounds.

What about premise (2)? Some people in the lecture suggested that there might be *circular causal sequences*: cases in which A causes B, B causes C, and C causes A.[2] If A, B and C were really *events* with specific distinct times (so that A occurred at $t$, B occurred at $t^*$, and C occurred at $t^{**}$), it would also have to follow that there would be some temporal circularity. I honestly have no idea at all whether this is plausible: perhaps some scientific results could push us in this direction, but I can't think of any clear examples. If one could, one could then question the argument that is built into (2): it wouldn't follow from the fact that causal sequences are temporally extended that nothing could be its own cause, because a temporal order could be circular. This seems mystifying to me, but I don't really have a neat argument against it. Due to the difficulty of finding clear counterexamples, I'd say (2) is the most promising premise in the argument.

### 2.3. *The Third Way*

Aquinas's third argument for God's existence is fraught with different problems. Recall it:

1. Some things are contingent beings: they might not have existed.

[2] Note that if we accepted the *transitivity* of the causal relation (i.e., that if A causes B and B causes C, then A causes C), this would also establish the possibility of self-causation, which we were doubting earlier.

2. For all contingent beings, there is some time at which they did not exist.
3. If everything were contingent, there would be a time at which nothing existed.
4. If there were a time at which nothing existed, then nothing would exist now.
5. Things do exist now.
6. So, there must be a necessarily existing thing.
7. That thing is God.

(1) is clearly true. (2), however, is *not* clearly true. Note that this general principle is false:

> **Modal Fallacy**. If it is contingent whether A is true, there is some time at which A is false.

Let A = the claim that I have two thumbs. It seems like this claim is contingent, since I might not have had two thumbs. But it doesn't follow that there's a time at which I'll lack a thumb. I certainly hope that's false, and will try, probably successfully, to see to it that it remains false. So, Modal Fallacy is false. But (2) is just an instance of this fallacy. So, there seems to be no reason to accept (2). Even more simply, (2) conflates *eternal* existence with *necessary* existence: perhaps there are elementary particles whose existence spans the whole duration of the Universe. I take it that these particles are contingent: they might not have existed. They're still *eternal*, though.

(3) is also highly problematic. Note that this general principle is false:

> **Quantification Fallacy**. If, for every X, there is some time at which X exists, then there is some time at which every X exists.

Everything does indeed exist at a time. But there are some things that don't exist at the same time: Socrates and me, for instance. This is why the Quantification Fallacy is a bad principle. Alas, claim (3) in Aquinas's argument is an instance of the Quantification Fallacy.

The rest of the argument is a bit less bad. (5) is clearly true. (6) is indeed a logical consequence of the claims that precede it. (7) is, alas, a product of the Additional Properties Fallacy, but Aquinas has already fallen foul of that many times, so big surprise. And (4) is not clearly false, though I also don't see that it's clearly true.

### 2.4. *The Fourth Way*

Let's turn to one last argument from Aquinas before turning to Paley's teleological argument:

1. Properties come in degrees.
2. To have a property to some degree is to approach something that has that property to the greatest degree.
3. Whatever is the greatest bearer of some property is the cause of all things that have that property to some degree.
4. So, there exists something that is the cause of all perfections and has all perfections to the greatest degree.
5. This thing is God, and so God exists.

Perhaps the most glaring new problem in this argument is the move from (3) to (4). There might, it seems, be *several distinct* things that are each respectively the cause of all things that have each distinct property to some degree. For instance, as Plato believed, there might be a Form of

Goodness that causes all things that are good to be good, and a Form of Oddness that causes all things that are odd to be odd. But the Form of Goodness needn't be the same as the Form of Oddness. So, why suppose, as the inference to (3) to (4) requires, that the perfections are all united in one perfect thing? There seems to be no reason.

A different problem with the argument is (2) is false. Some properties that come in degrees are *unbounded* in the sense that there *just isn't a greatest degree of that property*. Largeness may well be like this: while there may be a largest actual object, there is not a largest number: we can always conceive of larger numbers (including infinities, since there are infinite orders of them!). In cases where a property has no greatest degree, (2) will have to be false.

Finally, (3) seems bizarre at least if read literally (and, given the conclusion, I think it does have to be read literally). Suppose we have a small world with ten people. Half are thin. One of them is the thinnest. Is it necessarily the case that the thinnest is the sole parent of the other thin ones? No. Then in what sense could (3) literally be true? No sense, it seems.

**MEETINGS 2 and 3**

## 1. The Teleological Argument

Let's turn to a more interesting argument for God's existence: viz., the *teleological argument*, which is also often called the *argument from design*. James had you read Paley to see this argument, but the argument has been given by many people, including Aquinas (it seems to be part of his fifth way, which is why we're skipping to the somewhat clearer formulation in Paley).

The teleological argument can be formulated in several ways. Two are not deductive arguments. This doesn't automatically mean that they are bad arguments, though it does mean that they're formally invalid in the technical sense. It may, however, mean that they shouldn't be asserted from the armchair: for, to be good, a non-deductive argument has to rest on real correlations in our world, and such correlations are only really discoverable empirically. Such arguments are therefore subject to refutation on scientific grounds, and this is one reason why one doesn't often see such arguments being advanced by contemporary *philosophers* anymore: the arguments make assumptions that philosophers are not really equipped to argue for.

### 1.1. *The Analogical Reading*

One of the non-deductive readings casts the argument as an argument from analogy:

1. Organisms have functions and working parts directed at realizing these functions, and are like watches in this respect.
1. Watches have functions and working parts because they were designed.
2. Like effects tend to have like causes.
3. So, organisms have functions and working parts because they were designed.
4. Designers must be beings with minds.
5. So, some being with a mind designed human beings.
6. That's God, so God exists.

Notably, this argument seems to fall foul of the Additional Properties Fallacy just as much as the other arguments we saw in Aquinas. As James pointed out in the lecture, it's quite compatible with the premises of this argument that the minded being who supposedly created human beings is in fact a *shoddy* or indeed *evil* being. So, the move from (6) to (7) is invalid.

A new and distinctive problem with this argument owes largely to its non-deductive form. Note that (4) claims only that like effects *tend* to have like causes. They don't *always* have like causes.

That might be so in this case. How could it be so? One simple way is if there's a *different* explanation of the existence of the ostensibly function-possessing beings in the world like living organisms that doesn't appeal to a single conscious designer with all the perfections of God as "he" is traditionally understood. There are many possibilities: (i) living organisms came about just by chance, (ii) a large committee of less than fully perfect gods created the world, (iii) a shoddy, malevolent god is responsible for living organisms, and so on. Of course, some of these explanations might not seem to be so great. We have to postulate more entities if we opt for (ii), and so the single traditional God explanation seems *simpler* and thus better. Moreover, without some deeper account of the strictly physical, non-mental mechanisms that chancily produced living organisms, an appeal to (i) seems less attractive: *pure* chance is really no explanation at all. (But there is a popular chancy account on offer: natural selection.)

What else might be said against this reading of the teleological argument? One idea that someone suggested in the lecture that I think is serious is to deny premise (1), and claim that organisms and so on don't *really* have functions in the sense required for this argument to work.

How would this go? Someone in the lecture came up with a nice example which, when reflected on properly, answers my question. Note that flowing water can be used to generate power. This is a fact that comes from the intrinsic properties of water, which it has mind-independently. We can grant that there is *some* sense of 'function' on which we can then say: one function of flowing water is to generate power. Let's call this the *thin* sense of function, and define it as follows:

> *Function in the Thin Sense.* Y-ing is a function of X in the thin sense if and only if ("iff") X can be used to Y in virtue of its causal properties.

Now, let's ask a simple question: is there *any* reason to believe that function in the thin sense always requires explanation by a *designer*? I think not. Even if no mental beings ever existed in some world, it would be true to say of the objects in that world that they *could* be used for some purposes. So, function in the thin sense doesn't imply the actual existence of any designers. So, if *this* is the sense of 'function' that's relevant for the argument, the argument will fail: after all, we can understand how organisms could have functions in the thin sense without assuming any designer, and so (4) in the argument will be plausibly false.

'Function' could be used in a different, stronger sense. Notice that it's odd to say that a *purpose* of water is to generate power. Why? Well, because water doesn't *by itself* have any inherent purpose. Sure, water can do lots of things, and so has a function in a narrow sense. But it doesn't follow from this that it has a *purpose*.

Of course, some things do have purposes. It *is* true to say that watches have a purpose: namely, to tell time. So, we can understand a stronger sense of 'function':

> *Function in the Thick Sense.* Y-ing is a function of X in the thick sense iff one of X's purposes is to Y.

If *this* is the sense of 'function' that's used throughout the argument, then it has a different problem: namely, it's not at *all* clear that living organisms have *purposes* in addition to having functions in the *thin*, purely causal sense. To assume that they do in effect begs the question, since it is indeed quite plausible that everything with a function in the thick sense was designed. After all, the concept of *purpose* seems like a mind-dependent concept: things only have purposes in virtue of there being mental beings who *give them* these purposes.

So, there's a dilemma for the argument: if 'function' is understood in the thin sense in the argument, then (4) is arguably false, whereas if 'function' is understood in the thick sense in the argument, then (1) is arguably false. Either way, the argument would be unsound.

### 1.2. *The Abductive Reading*

There is a different non-deductive way of reading the argument, but I think it suffers from the same problem. Recall James's abductive reading of the argument:

1. Some natural things have functionality.

2. The best explanation of this is intelligent design.
3. So, some natural things are the product of intelligent design.

This argument is open to precisely the same kind of dilemma as the last. If 'function' is understood in the thin sense, then (2) is false. After all, it would be simpler and indeed more natural to understand the thin functionality of some natural phenomena without appealing to any mental entity: after all, things can easily have functionality in this sense without being designed. Indeed, this explanation would be simpler in two senses: we'd have to posit fewer *numbers* of things (since God would be an extra being), and we'd also have to posit fewer *kinds* of things (since God is a supernatural being, but we can explain function in the thin sense in strictly naturalistic terms).

What if 'function' is understood in the thick sense? Well, then it's not even clear if (1) is true. To just assume that it is would seem to beg the question, since it may indeed be *definitional* that things have function in the thick sense only if some entity *gave* them that function. But we can't *just observe* that there was a designer. And, in any case, that's supposed to be the conclusion of this argument, not one of the hidden premises of it!

### 1.3. *The Deductive Reading*

Alas, precisely the same problem even more obviously destroys the last reading of the teleological argument, which is the following deductive reading:

1. Some natural things have functionality.
2. As a matter of conceptual truth, things with functionality are products of intelligent design.
3. So, some natural things are the product of intelligent design.

The same dilemma shows up here. If 'functionality' is understood thinly, (2) is clearly false, whereas if it's understood thickly, (1) is doubtful and asserting it without argument would seem to beg the very question at issue.

### 1.4. *Final Point about Function in the Thick Sense*

There is one last point worth making that has bearing on all three readings of the argument. So far, I've been running my dilemma by saying that if 'function' is understood thickly in any of the readings of the argument, then these readings beg the question, since the question at issue is whether any natural phenomena were products of intelligent design. We can't just assume that that's true, which we'd have to do if 'function' were understood thickly.

But more can be said. I think we do want to say that human beings have some sort of purpose. Is that enough to get the teleological argument off the ground? No, it's not, because it's quite coherent to suppose that humans *give themselves their own purposes*. *We* get to decide what our purpose is, in part because we're *free*. If that's true, then *even if* some natural phenomena (viz., human beings) have functions in the thick sense, the explanation of that fact needn't appeal to any further phenomena: we can just appeal to the self-determining nature of human beings to explain how human beings can have purposes, and build things up from there.

This is a natural idea, and one that's at the core of some famous movements in Continental philosophy like *existentialism*: the autonomy of human beings is the only source we need to understand the sense in which human life has a purpose. Perhaps this is unsettling: it is then up

to us what we make of ourselves, and there is no third party that decides it for us. But what's unsettling to some might be liberating to others! And this strikes me as quite liberating.

## 2. Some Points about "Subjective Truth" and Related Issues

Several times in the lecture people have appealed to the idea that certain truths (or perhaps even all truths) are "subjective". This is an idea that can be understood in many different ways, some of which are trivial, and some of which are much more substantive but also much less obvious. It is important not to appeal to this idea unless one has a clear sense of what one intends by it.

In appealing to this idea, people often say things like:

> *The Quick Subjectivist Assertion*: "What's true for some people isn't true for others."

This assertion communicates something of value, but I don't think it communicates something immediately plausible when read *literally*. So let's focus first on some less literal ways of reading it that are immediately plausible:

1. People have different belief systems, and so different things will *seem* to be true to different people.

2. People can have irreconcilably different evidence, and so different things will be irreconcilably rational to believe to be true for different people.

I suspect that (1) is what people really want to communicate when they make the Quick Subjectivist Assertion. But (1) is compatible with there being mind-independent, universal, objective truths. So is (2). So if that's all that one wants to communicate by making the Quick Subjectivist Assertion, one doesn't literally believe that truth is subjective. The things one believes are subjective are things *everyone* can agree are subjective: namely, beliefs or evidence.

I should soften the tone here just a bit. While I think (1) is a trivial truth that offers no cause for excitement, (2) is a much less trivial truth *even though* it is compatible with objective truth across the board. The fact that (2) is true should encourage us to adopt an attitude of humility in engaging in argumentation about many matters. This is an attitude that some people incorrectly believe is incompatible with belief in objective truth across the board. Sometimes one hears opponents of objective truth charging those of us who believe in it with arrogance. They think the claim that some truths are objective sounds like some sort of cultural imperialism, and get angry about this for (otherwise completely sensible) political reasons. It's crucial to see that this claim entails nothing arrogant or imperialistic, because it's compatible with the truth of (2).

If people can have irreconcilably different evidence for believing things – which is clear if *intuitions* count as evidence – they could reasonably disagree with each other and be unable to resolve this disagreement. And if that's right, then there are situations in which neither of them could "win the argument", since they would properly continue to disagree at a bedrock level. This is a fact that demands us to be humble in engaging in argumentation: we ought to be open to the fact that people's intuitions may just fundamentally differ from ours. If so, then although it's still quite true that, insofar as we really disagree, only one of us can be right, it may be impossible for us to determine which of us is right. That's a very serious point.

Now, could *truth* really be subjective, rather than just *believed truth* or *rationally believed truth*? To answer this, it's useful to introduce the concept of a *proposition*. Suppose Jim and Jones both say:

Sentence Type I: "I am bald"

And suppose that Jim is bald and Jones isn't. Something is thus false when Jim says Sentence Type I, and something is true when Jones says Sentence Type I. But if they really are *saying the same thing*, then one and the same thing is both true and false. That is a consequence it would be nice to avoid, since it looks incoherent. What's really the case is that they *aren't saying the same thing*, though they are *using the same words*. To understand how this is possible, we need a concept of what's expressed by a sentence that isn't itself a linguistic entity. We call this a *proposition*: Jim and Jones assert different propositions by saying Sentence Type I.

So, we can talk about two things as bearers of truth and falsity: sentences and propositions. Sentences are at best *secondary* bearers of truth and falsity, since they are only true or false in virtue of expressing propositions that are true or false. Here is one view one could take:

> *Trivial Subjectivism about Sentence Type Truth*: Some sentence types express different propositions depending on who utters them, which needn't all be true.

This view is clearly true: the case involving the two utterers of Sentence Type I shows it to be true. There is a less trivial kind of claim one could make that strengthens this claim:

> *(Schematic) Revised Subjectivism about Sentence Type Truth*: Every sentence type S about a certain subject matter M is such that, for possible speakers A and B, if A and B assert S, they express different propositions which needn't both be true. In each case, neither proposition is "privileged" in being the *right meaning* for S.

This is the kind of view I think people really want to express when they say that certain truths are merely subjective and intend something literal by saying it. Now consider:

> *Revised Subjectivism about Sentence Type Truth in the Ethical Domain*. For every sentence type S about ethical matters, there are possible speakers A and B, such that if A and B assert S, they express different propositions which needn't both be true. In each case, neither proposition is "privileged" in being the *right meaning* of S.

On this kind of view, there might be two people who say, "Killing is wrong", where one of them asserts a true proposition, and another asserts a false proposition, and where neither of them is mis-using words, and neither proposition is the privileged meaning of the sentence. If this kind of view were correct, there would be a deep sense in which ethical issues are subjective.

This kind of view, however, is not a view that it is easy to argue for. One can't just assume it out of nowhere, and claims like (1) and (2) above do not provide arguments for it. The fact that people have different beliefs or have different evidence regarding moral matters does not entail that they assert different propositions when they say things like "Killing is wrong", and that neither proposition is the privileged meaning of the sentence. So, it certainly won't do to appeal to this hugely controversial idea out of nowhere in trying to resolve a philosophical problem.

Revised Subjectivism about Sentence Type Truth in a domain is rarely plausible. We usually think people *really disagree* with each other when they argue about whether abortion is permissible. Why else would they care so much and get so impassioned? But if Revised Subjectivism on ethical matters were true, this would be false: they're just expressing different

propositions and neither is the privileged meaning of the sentence, so both could be right and there's nothing to argue about. They could resolve their dispute by saying: "We're not even talking about the same thing, so there's no sense in us arguing as if we were at odds with each other. Let's just be friends!" Since that's implausible, I think we shouldn't accept this view.

There is a different nontrivial way in which certain truths could be subjective that is more tenably exemplified. I doubt it will clearly help people with subjectivist sentiments about ethics or religion, but it does capture a deep sense in which certain matters are *mind-dependent*.

This way is brought out most easily by considering what Locke calls *secondary qualities*, such as colors and smells. I take it that we want to say that grass is green. If *anything* is an uncontroversial empirical claim, this is. But notice that whether grass is green would seem to depend in *some* way on facts about us, like the construction of our eyes and our perceptual processing mechanisms. If our brains were different, grass could have looked very different. If the lenses in our eyes were different, grass could have looked very different. Given suitable tinkering with our eyes and our brains, it would be easy to make grass appear to have precisely the color that stop signs now appear to us to have. Indeed, we all *could* have been born this way.

If that's right, then in what sense is grass *really* green? We can, I think, agree that it's true that grass is green. If anything is true, that claim is. But once we start reflecting on the vast range of different ways in which our mechanisms of perception could have been constructed (or on the *actually* vastly different ways in which the perceptual faculties of different animals *are* constructed), the fact that this claim is true seems like a red herring. *This* true claim seems not to do very much to reflect the real nature of the external world. It seems to reveal more about *us* than about the world. If that's so, then it looks like there's a deeper sense in which truths such as that expressed by "Grass is green" are subjective in a way that should detract from our interest in them if we want to know what the world is *really like*.

Let's summarize this view about color truths as follows:

> *Lockean Subjectivism about Color Truths*: Although some claims about whether external objects have color are true, their truth reveals little or nothing about the *real nature* of those objects. And given conceivable (or indeed actual) perceptual variability, the truth of these claims is clearly mind-dependent.

This type of view generalizes fairly dramatically. Almost everything that we care about in our ordinary activities is mind-dependent in a significant way: how things taste, feel, sound and smell are *all* subject to the same vast possible perceptual variability, and so Lockean Subjectivism will be true of most of the truths that we observe. If that's right, then although there surely is a mind-independent world, we may have very little by way of *direct* or *revelatory access* to it. And this, it seems, is a thought that should encourage even more humility in us than (2) above.

(It's worth mentioning the views of the 18th Century philosopher Immanuel Kant here, since we won't otherwise be getting much of a chance to discuss them. Kant thought, plausibly, that human beings are hard-wired to interpret experience by applying certain conceptual schemes and categories. He called the empirical world as it is filtered through our conceptual schemes and categories the "phenomenal" world. Kant was plausibly agnostic about whether the conceptual schemes and categories that we're hard-wired to apply to the world do anything to reflect its real nature. So he distinguished the "phenomenal" world from what he called the "noumenal" world, which is just the way the world is independently of the concepts and categories we are irresistibly predisposed to use.

Arguably, reflection on the way in which Lockean subjectivism generalizes encourages this type of view: if our mind paints onto the world colors, smells, tastes, textures, shapes, and other features that it doesn't *really* have, but we can't get in touch with it without viewing it as having these properties, it seems like the right conclusion to draw might be that *we don't observe the world at all*: we're stuck behind a veil of appearances ("phenomena") that we can never really penetrate.)

The Lockean view could generalize to other cases where people have wanted to insist on unavoidable subjective taint. I doubt that it works for claims about God, but it may work for claims about ethical matters. Recall how, in our first meeting, I noted that most argumentation in ethics often (and perhaps always) comes down to bare appeals to moral intuition. One *might* think that creatures with different physical and psychological constitutions could have different moral intuitions. Suppose we did in fact have very different moral intuitions because we were very differently physically/psychologically constituted. If that were so, it would *at least* be clear that the ethical theories on which we would converge at the end of the day would be different.

Of course, this point may seem on first glance compatible with our just being *wrong*: we can after all imagine a harsh society of nothing but cold-blooded sociopaths who don't find it compelling that killing is wrong, and they would arrive at very different ethical theories at the end of the day, and we would reasonably regard them as mistaken. This, however, is misleading, since here it wouldn't be true that *everyone's* moral intuitions were the same and also similarly mistaken: *we* would still be right, one might say, in reacting negatively to this group of people. It's the idea of *everyone's* moral intuitions being *massively* unreliable in the sense that they are *always* mistaken that seems hard to swallow. How could it be the case that a sentence of the form "A-ing is wrong" is true when *everyone*, after indefinite amounts of reflection, continued to have the intuition that such a sentence is false? This idea seems nearly incoherent: *moral truths aren't inaccessible in this way*.

But if that's right, then it seems not crazy to think that some moral truths could be mind-dependent in the same sense in which claims about colors and smells are mind-dependent. After all, it's implausible in just the same way to suppose that *everyone* in some possible world could be *misrepresenting* grass if they had the kind of qualitative color experience we have when we look at stop signs when looking at grass.

Could *all* truths be subjective, not just *some*? I think a positive answer to this question does not express a coherent view. We can refute such an answer by a *reductio ad absurdum*:

1. Assume for the sake of argument that all truths are subjective.
2. This putative truth – i.e., that all truths are subjective – is itself subjective if it's true.
3. This means either (i) that two people could assert this claim, it would express different propositions in their mouths which aren't both true, and neither proposition would be the privileged meaning of the claim, or (ii) it reveals nothing about the real nature of the world.
4. If it means (i), then someone asserts a falsehood when he asserts that all truths are subjective.
5. If a view entails that someone who asserts it asserts something false and isn't asserting a proposition that isn't the right meaning of that sentence, that view should be rejected.
6. So, if the claim means (i), it should be rejected.
7. If the claim means (ii), then (A) there would be a real nature of the world that the claim didn't capture, and this real nature would *ipso facto* not be subjective.

8. (A) also implies that it would be true that this real nature would *ipso facto* not be subjective.
9. So, (A) couldn't itself be merely subjectively true, since it would then be false: after all, for it to be true, it has to be true that there is a real nature that the world has that isn't subjective.
10. So, if the claim means (ii), it entails its own falsity.
11. So, either (1) should be rejected or it entails its own falsity.
12. This contradicts (1).
13. So, (1) is false.

This is a valid argument that shows that the claim that all truths are subjective is self-refuting, and I think it's quite a serious argument. I don't doubt that there are replies to it, but the burden of proof certainly seems to be on the defender of the view it attacks. As Chris Swoyer expresses the thought behind this argument in his entry on relativism in the Stanford Encyclopedia of Philosophy: "Relativists always face the occupational hazard of sawing off the limb they're sitting on, but with [the claim that all truths are subjective] they seem to cut down the whole tree."[3]

So, the upshot of all this discussion is that there may be some sense in which *some* truths are subjective, but if it's going to be a nontrivial sense, one is going to have to engage in some serious argument for it, and one is *not* going to be able to establish that all truths are subjective without ruining one's own dialectical standing.

## 3. Anselm's Ontological Argument

Let's turn at last to what is perhaps the most intractable and interesting argument for God's existence. Anselm's argument, as you'll recall, is a *reductio ad absurdum* that goes like this:

(1) Assume for the sake of argument that God does not exist in reality.
(2) By definition, God is that than which none greater can be conceived.
(3) God exists in the understanding.
(4) So, God exists in the understanding but not in reality. (From (1) and (3))
(5) If a thing exists in the understanding, it is conceivable that it also exists in reality.
(6) A thing that exists in understanding and in reality is greater than a thing that exists in the understanding alone.
(7) So, a being greater than God is conceivable. (From (4), (5) and (6))
(8) (7) and (2) contradict each other, and so an absurdity follows from (1).
(9) So, by *reductio ad absurdum*, we can conclude that God exists in reality.

Before we try to pick apart this argument, it's worth noting how much better it is than most of the arguments we've seen so far for God's existence.

One clear virtue of the argument is that it's formally valid: it's a *reductio ad absurdum*, and all arguments of this form are formally valid. Another clear virtue of the argument is that it doesn't seem to fall foul of the Additional Properties Fallacy. This kind of argument can indeed be used to show that God has all the perfections he is often assumed to have, since if he lacked these perfections he would be a being than which a greater being with those perfections could be conceived. So, when we arrive at the conclusion of this argument, we get quite a lot.
A final virtue of the argument is that it's hard at first to see where it goes wrong! This argument has no *obvious* fallacies or false premises. All the steps have some initial intuitive plausibility. (2)

[3] http://plato.stanford.edu/entries/relativism/#5.9

seems like a fair way to capture what people standardly believe about God's perfect nature. (3) seems trivial at first glance (though not, as I'll be suggesting, at second glance!): surely we have a *concept* of God, and surely that's enough for God to "exist in the understanding" in the sense at issue in this argument. (1) is just an assumption for *reductio*, and so it's not part of the argument that can be questioned, since it's the very assumption whose ability to entail an absurdity gives us a reason to reject it in the conclusion and hence conclude that God exists. (4), (7) and (8) are consequences of the preceding claims. (6) seems obvious. So, really the only premise that isn't *prima facie* obvious is (5), but even it isn't *implausible*: lots of people in the history of philosophy have assumed that understanding something requires conceiving of it as possibly real.

Still, further reflection reveals some subtle loose ends that, when pulled at, untie the whole argument and threaten to collapse it. What I think is the central problem centers on the relationship between claims (3) and (5). To bring this out, let's ask why (3) seems plausible.

Insofar as we do understand what God would be if he existed, what does this understanding amount to? As far as I can see, we can't understand what it would be for there to be God by *imagination*: I can't *visualize* God, or picture in my head what it would be for God to exist. So how do I understand what it would be for there to be God? Part of it is just by understanding the concepts that are used to identify God: he's a being will all sorts of perfections, like benevolence, omnipotence, omniscience, and so on. Since I understand what it is for someone to be benevolent, powerful, and knowing, and (maybe!) to have these properties to the greatest degree, I can *ipso facto* understand the conjunction of all these properties. Since whatever instantiates that conjunction is God, it seems that if I understand him at all, this is how I do it.

Let's give a name to this kind of understanding:

> *Understanding by Concept Composition* ("Understanding$_{CC}$"): A person S *understands X by concept composition* (or "understands$_{CC}$ X") if and only if S understands a range of concepts that can jointly be used to define what it is to be X.

What I've been saying is that if we understand what it would be for there to be God, we understand it by concept composition: by joining together a bunch of properties that God is supposed to uniquely have.

But now we've got a source for trouble in the argument. Notice that if the argument is to remain valid, there can't be an equivocation on the meaning of 'understanding' in it, so that "understands" is used in one sense in one premise and in another sense in another premise. So, if we have to understand 'understanding' as "understanding by concept composition" in premise (3) – which I'm suggesting we must – we also have to understand 'understanding' in this way in premise (5). If that's so, then premise (5) can be stated more explicitly as follows:

> (5-Unpacked) If we understand$_{CC}$ something, then it is conceivable that it also exists in reality.

But *this* claim is false. Understanding by concept composition is a quite liberal thing: if I understand two concepts, then I *ipso facto* understand$_{CC}$ anything built out of those concepts. Well, I understand the concepts of roundness and squareness. So, I understand$_{CC}$ what it would be for something to be a round square. But clearly a round square cannot exist in reality, and we cannot conceive of it existing in reality. So (5-Unpacked) is false.

If that's right, the argument is unsound. Of course, it's open to Anselm to rely on a less liberal notion of understanding. But if he does, it's not obvious that (3) will be true. After all, recall how we started. I started by suggesting that there are various ways of understanding 'understanding' on which (3) couldn't be true. One way to understand something is to be able to competently imagine what it would like for that thing to exist. We can't, I suggested, understand God in this sense. Indeed, it really seems like our only access to the concept of God is concept-compositional: we understand God by understanding various properties he's supposed to have, and by trying to put those properties together in our heads.

So, to put my criticism a bit more accurately, there's a dilemma for this argument. If 'understanding' means something as weak as *understanding*$_{CC}$ – which it seems it must – then (5) will be falsified by the case of the round square. If, on the other hand, 'understanding' means something stronger, then (3) will no longer be plausible, since we *don't* seem to understand God in anything but a concept-compositional way.

Are there other problems with the argument? None, I think, that are quite as clear as this one. The only other spots in the argument that we could *try* to question directly are (2), (3), (5) and (6), since the rest of the premises either follow logically from other premises or are structural assumptions needed for the *reductio ad absurdum* to work validly (e.g., premise (1)).

The only argument I've heard against (6) doesn't work. Someone in the lecture suggested that our ideas of things often turn out to be better than the things themselves. The only sense in which this claim communicates something true is a sense that's irrelevant. To see this, suppose I imagine having a blind date. I imagine it turning out really good: my date is smart, funny, gorgeous, and really digs me. Then I have a blind date, and it turns out quite differently from what I imagined – indeed, much worse. Of course, if what I imagined had been actualized, *that* would have been better. The problem was just that what I imagined *wasn't* actualized: my blind date ended up being *different* (and worse) than I imagined. For it to be the case that our idea of something is better than the thing itself, the idea needs to actually be an idea *of that very same thing*. But my idea wasn't really about the blind date *I ended up having*: it was about a different date which, if actualized, would have been better. So, we can't attack (6) on this basis.

So, the only other premise we could attack is (2). (2) was supposed to be a definition, so it seems hard to attack it in any direct way. One thing we could do is to argue that the definition is incoherent. And this, I think, isn't a completely crazy move. Remember what God was supposed to be: he was supposed to be a being with every perfection to the *greatest degree*. But some perfections may be *unbounded*, and not have a greatest degree. I can always imagine a better world: a world with more happy people in it, more beauty, and so on. I can then also imagine a being that creates such a world. As long as the sequence of worlds is infinite, the sequence of creators will also be infinite. And *ipso facto* as long as there isn't a greatest world (which by assumption there couldn't be, since I could keep imagining better), there won't be a greatest creator. So, if some perfections are unbounded, the very concept of a being than which none greater can be conceived may be empty. This, I think, is an interesting further line of criticism, though I'm not sure it's as decisive as the dilemma I initially sketched.

**MEETING 4**

## 1. The Problem of Evil

The world seems to have a lot of evil in it. This evil comes in two forms: some is generated by the voluntary acts of conscious beings (e.g., the Holocaust), and some occurs in nature (e.g., tsunamis and diseases). On the face of it, the fact that evils of either kind exist conflicts with the thought that there exists a God who is omnipotent, omniscient and benevolent. And so the *Problem of Evil* arises for believers in such a God – i.e., the problem of resolving this conflict.

The problem can be expressed in a couple of ways. The strongest is the *Conceptual Problem of Evil*: the existence of evils of either kind seems *conceptually inconsistent* with the existence of such a God. This problem is made poignant by the following brief deductively valid argument for atheism:

> *Argument for Atheism from the Conceptual Problem of Evil*
>
> (1) By definition, if God exists, he is omnipotent, benevolent and omniscient.
> (2) Alas, evil exists.
> (3) But, by (1), if God exists, he must have the power to eliminate this evil, he must know that it exists, and he must have the desire to eliminate it. In short, he could and should have gotten rid of this evil if he exists and (1) is true of him.
> (4) So, God doesn't exist.

The scope of this argument is not as grand as one might like it to be. Some people might be happy to believe in supernatural beings that are less perfect than the God some religious traditions recognize when they endorse (1). So, strictly speaking, this isn't really an argument against *any conceivable god*: it's just an argument against a fully perfect one, at least assuming omnipotence, benevolence and omniscience are indeed perfections. Still, enough people have believed in such a being that this argument raises a serious problem worth addressing.

One might have thought that it's not so clear that the God accepted by the most widespread religious traditions really *is* benevolent. He is, after all, described as *wrathful* in the scripture of those traditions. But this thought is premature: it's not as if God just randomly gets ticked off and decides, for no reason other than to express his anger, to do unpleasant things. God mainly gets ticked off at wrongdoing and sin. It's coherent to suppose that a benevolent being could get angry about stuff like this. Surely a benevolent being could want to punish wrongdoing: if my mom, a paragon of benevolence, had the legal authority to sentence some serial killers to a long stay in prison, and she did so, I don't think we'd want to say that she was less benevolent for it. The same goes for God. (Of course, there are some related worries: eternal damnation seems like a gratuitously stiff penalty for some of the sins on which God is inclined to use it according to the scripture of some religious traditions. Here someone might reasonably doubt whether such a God is fully benevolent.)

So, in any case, let's focus on the concern that the Conceptual Problem of Evil seems to raise for those widespread religious traditions that do hold that God is benevolent, omnipotent and omniscient. Should we worry about the Argument for Atheism from the Conceptual Problem of Evil if we are members of these traditions?

Well, it is a very tough argument: the only premise that seems moderately negotiable is (3). But why would a benevolent God *not* have the desire to eliminate the evils of the world?

Theists do have responses to this question. One is to appeal to the *value of free will*. Much of the evil in the world is attributable to the voluntary acts of conscious beings, including but not limited to those of humans (perhaps some non-human animals also have free will and use it for evil). One might think that free will was a *gift*, and that its value is great enough to outweigh the bad things that it sometimes is used to produce. If so, then, on balance, it may be best for the world to be as it is.

One problem for this response to the argument is that it does nothing to address natural evils. One is going to have to tell a pretty farfetched tale to make it plausible that all natural evils are attributable to voluntary acts of conscious beings. Maybe *some* are: global warming, for instance. But there were plenty of earthquakes, hurricanes and diseases that killed perfectly innocent people well before we started that mess. Perhaps one could claim that these natural evils are due to the agency of other lesser supernatural beings (e.g., demons) that God also created. If free will really were so valuable that it could outweigh the bads perpetrated by human beings, one might think it could also be valuable enough here. *Maybe* this is right: it isn't *conceptually incoherent*, and the problem here is supposed to be a conceptual one. But this response isn't sufficient to undermine the Problem of Evil in all its forms: for it can take an inductive form as well as a conceptual form, and it just seems *unlikely* that God would have created such demons when he could have created benevolent supernatural beings or lots more happy human beings instead.

But there are even more basic problems with the free will defense. One of them is that it seems like God could have just created *really virtuous free beings*. Crucially, there's nothing incoherent about the idea that beings could be *free but have constraints on their behavior that are due to their characters or personality traits*. To see this, consider the fact that most of us can't even *conceive* of trying to kill another innocent, healthy person that we like. If you put a knife in my hand and asked me to stab you to death, and it was clear that this wouldn't amount to euthanasia or rightful punishment, *I just couldn't bring myself to do it*. It literally isn't even possible for me to do that in any ordinary case.

And that's thanks to my personality. So, when constraints on behavioral options are imposed by the intrinsic nature of a person's character, we don't think these constraints make that person lack freedom in any interesting sense. And so there are limitations on my agency that don't make me lack free will. Would I have been *freer* in any interesting sense if I could make myself stab you to death on command? *Maybe* (though I doubt I'd even be *me* in that case, so I'm not so sure this is conceivable), but I'm *still free* as it stands, and it looks like removing the constraints on my freedom that are imposed by the inclinations against random killing produced by my minimally virtuous character wouldn't make me freer in a way that would be *more valuable*.

But if that's right, then why couldn't God have made it a law of nature that only virtuous, kindly beings with all the good constraints on character that various saints in history have had could survive in this world? These beings would still be free – unless free will is simply impossible, which would ruin this defense of theism! For, as we've seen, if constraints on behavior just grow out of the virtues of your personality, we don't think these constraints deprive you of freedom. The same holds for the far more virtuous beings I'm imagining.

Now, one could try to say: but God *made* these beings have this character, and so in some sense their acts are *determined by God*. But this objection fails: it overgeneralizes in an obviously embarrassing way. *Precisely* the same is true of me if God exists: God could have arranged nature in a way that gave me a different character, but he didn't, and so his acts determined my character in just the same way. So, if we continue to think I'm free, there's no further reason

why we shouldn't think these perfectly virtuous beings are free. So, why didn't God make the world have them instead of all the bad people? The free will defense seems to offer no clear answer to this question. Without an answer, the Argument for Atheism from the Conceptual Problem of Evil would seem to go through.

There are other problems with the free will defense. One is that free will just cannot have the kind of value it needs to have for the defense to get off the ground. There have been spectacularly terrible people in this world (e.g., Hitler). Why didn't God do something to prevent them from coming into existence? If he's omnipotent, he could have, and if he's omniscient, he could have seen this coming (at least on probabilistic grounds, given knowledge of their character). It looks like the proponent of the exceptional value of free will has to say that the value of Hitler's free will *by itself* was strong enough to give God a decisive reason not to prevent him from coming into existence.

Besides being crazy on its face, this claim has odd implications for what *we* ought to do. The same value that by this evaluative hypothesis gave God a reason for action would give us a reason for action. If Hitler's mother, when four months pregnant, could have foreseen what he was going to do, it looks like it follows that she should not have gotten an abortion *simply because* the value bestowed upon the world by his free will would outweigh the evils he'd perpetrate. This is absurd. And it remains absurd even if you think abortion is sometimes wrong for *other* reasons: surely the fact that Hitler was going to be a further free being couldn't *by itself* have given his mother a decisive reason against abortion if she knew what he was going to do.

There are related problems. Suppose genetic screening reveals that if a certain couple tries to conceive a child, it is going to have a horrifying and completely incurable disease that's nevertheless compatible with free will – say, a severe nerve disorder that will leave the child in radically unbearable, unremitting pain for his entire life, but that would still leave the child free enough to develop and act and do some things. If free will has the kind of value that the free will defense suggests it does, it follows that the prospective parents in this case could reason like so: "Although we will bring a miserable being into existence that will suffer egregious and unpreventable pain, we will also be bringing a *free* being into the world. Since that's enough value to outweigh the disvalue of its suffering, we have, on balance, decisive reasons to have a child." This is not good reasoning.

We don't even have to think about cases involving the beginnings of existence to see this point. Should we unlock all the serial killers in prison and let them roam free and choose as they like? No. But we are constraining their freedom. The free will defense requires that unconstrained freedom has a huge amount of value that could on balance outweigh all the evils they would perpetrate. If so, it looks like we have a reason to free these serial killers. But this isn't so.

So, it is hard to see how free will could be as great a thing as the free will defense requires. Moreover, the free will defense just doesn't seem equipped to explain natural evils. So, what else might be said against the Argument for Atheism from the Conceptual Problem of Evil?

The only other clear idea that might seem to help is one that James usefully called the "character building response". On this line, enduring great evil can make us appreciate the good in our lives more, and can even contribute to the value that life has as a whole.

This idea is definitely plausible in a range of cases. But there are other cases sufficient to raise the Problem of Evil where it doesn't seem to work. In general, people cannot remember their

days as infants, though infants clearly are conscious beings. Accordingly, if someone suffered egregiously but only as an infant, this person would not probably remember this, and would not be able to see, later in life, how much better his life is now than when he was suffering as an infant. It's hard to see how, in this kind of case, there could be the kind of character building that the response we're considering needs. In what way did it make someone's life more valuable as a whole to suffer as an infant, or perhaps even, in cases of great evil, to be tortured as an infant? I just don't see how this is plausible. Moreover, I believe that the suffering of nonhuman animals is a natural evil. The world would be better if nonhuman animals suffered less. But most nonhuman animals just don't have the kind of psychological complexity that would allow them to better appreciate the goods on the basis of comparison with the bads. So, this story will be insufficient to explain why these sorts of evils lead on balance to greater good.

It also seems like there is a threshold beyond which experiencing greater suffering doesn't build one's character in an increasingly valuable way. Would people who were cured of terrible diseases be rational to wish *ex post facto* that they had suffered from a *worse* disease, because it would have been *even more character-building*? This seems implausible. Yet suffering frequently passes the threshold that seems to contribute to valuable kind of character building to which this response points. So it looks like there is going to be some really terrible evils that this response cannot explain.

The free will defense and the character building response are the only remotely plausible responses to the Conceptual Problem of Evil of which I'm aware. It is, then, a very burdensome problem for many theists, since these responses do not seem sufficient to solve it.

## 2. Pascal's Wager

Seeing how hard it is to provide *epistemic* reasons for believing in God's existence (i.e., reasons that favor of the *truth* of this belief), one might be inclined to start looking for pragmatic reasons. This is Pascal's tack: he gives a *decision-theoretic argument* for believing that God exists.

The core idea of decision theory is that it is most rational to do what would have the greatest expected utility. How does one compute the expected utility of making a decision? One figures out all the relevant ways the world could be if one made that decision, assigns subjective probabilities and utilities (i.e., degrees of perceived value) to these possibilities, takes the product of each probability and associated utility, and then adds up all of these numbers. Here is a simple example. Suppose you could pay one dollar to play a coin-flipping game. If the coin comes up heads, you get 5 dollars. If it comes up tails, you must pay another dollar. Your task is to decide whether to play this game. You know the coin is fair. Should you play? We can construct the following decision matrix to get the answer:

| | **Heads** | **Tails** | **Expected Utility** |
|---|---|---|---|
| **Play** | (5-1)(.5) = 2 | (-1-1)(.5) = -1 | 1 |
| **Don't Play** | 0 | 0 | 0 |

Since the expected utility of playing is greater than that of not playing, you'd be rational to play.

Pascal has a table just like this that is constructed from the following assumptions. He thinks that if you believe in God, and God exists, you'll other things being equal go to heaven, and that's an outcome with *infinite positive utility*; if God doesn't exist, you'll lose nothing, and have a life worth living anyway. If you don't believe in God, and he exists, you'll other things being

equal go to hell, and that's an outcome with *infinite negative utility*; if God doesn't exist, you'll lose nothing, and have a life worth living anyway. So, it would be more rational to believe in God than not to:

| | **God exists** | **God doesn't** | **Expected Utility** |
|---|---|---|---|
| **Believe in God** | $(+\infty)$(s.pr. $>$ $0 < 1) = +\infty$ | $(x > 0)$(s.pr. $> 0 < 1) > 0 < \infty$ | $+\infty$ |
| **Don't believe in God** | $(-\infty)$(s.pr. $>$ $0 < 1) = -\infty$ | $(x > 0)$(s.pr. $> 0 < 1) > 0 < \infty$ | $-\infty$ |

Although the argument is simple, it rests on a bunch of assumptions, some of which aren't totally clear. I'm no expert on the religions of the world, but I'm not sure that there is decisive scriptural evidence in all the major religious traditions for thinking that God punishes *agnosticism* with an afterlife in hell. And being an agnostic is one way in which you could refrain from believing in God. If this is unclear, then the table has to be made more complicated.

The argument also assumes that it would be irrational to assign a subjective probability of zero to the proposition that God exists. This is close – but not close enough! – to a standard assumption that gets made in the theory of rational subjective probability.

Call probabilities 1 and 0 *extremal probabilities*, because they are the upper and lower bounds of the probability space. It is standardly thought that assigning extremal subjective probabilities is permissible only in a couple of cases: you can assign probability 1 to *logical*, *conceptual*, and *mathematical truths* (e.g., "If P, then P", "squares have four sides", "2 + 2 = 4") and, correspondingly, probability 0 to logical, conceptual, and mathematical falsehoods (e.g., "P and not P", "squares have only three sides", "2 + 2 = 3"), and you can assign probability 1 to propositions that express your *evidence* or *what follows deductively from your evidence*, and hence probability 0 to propositions that deductively conflict with your evidence. Now, evidence is normally taken to include things that we can know by direct observation: I can assign probability one to the claim that there is a table in this room, because I can see that there is a table. And if you can assign probability 1 to claims in your evidence base, you can also assign probability 0 to anything that your evidence rules out: so, I can assign probability 0 to the proposition that there is not a table in this room, because my evidence entails that this is false.

But now suppose that you are convinced by the Argument for Atheism from the Conceptual Problem of Evil. If you are, you think that it follows from *a fact that you can observe* – namely, that there is suffering – and *a conceptual truth that you know* – namely, that if God exists, he is benevolent, omniscient and omnipotent – that God does not exist. And so the claim that God doesn't exist could, it seems, deductively follow from your evidence. If you can assign probability 1 to your evidence, you can assign probability 1 to what follows deductively from your evidence: namely, that God doesn't exist. And if that's so, then you can also assign probability 0 to what is inconsistent with that claim – namely, that God does exist.

So, it's far from clear to me that, if we actually take deductive arguments against God's existence seriously, we are not rationally entitled to assign probability 1 to the claim that God doesn't exist. And that would ruin the argument, since then we'd be multiplying the negative infinity in the lower left-hand box by zero, which would give us zero overall in that box.

This, however, is a fairly extreme kind of criticism, and I don't think it's decisive. Mainly I'm just pointing to a different controversial assumption that James didn't, I think, do enough to question.

How else could this argument fail? Many people are inclined to insist that this argument rests on a bad analogy. They say: "You can't just get yourself to believe things at will. So, beliefs are not objects of choice. And yet this argument presupposes that they are. So, this argument presupposes something false." Call this the *Objection from Doxastic Involuntarism*.

Like James, I don't think that this is the best objection to Pascal's wager. There are two related reasons for this. To see the first, notice that there are two different ways in which one can have control over an outcome: one can have the ability to make the outcome obtain *at will*, and one can have the ability to do at will some different things that eventually lead to the outcome. I can't make my life be a really excellent life at will (i.e., like I can raise my arm at will), though that would be really nice. Should I think I have *no control* over whether my life is an excellent life? No, because there are lots of other things I can do at all that can eventually contribute to whether my life ends up happy. Let's call this weaker kind of control *indirect control*.

It seems clear that there are lots of things we could do that could affect whether or not we believe in God. As James was noting, most people tend to be heavily influenced by the beliefs of the people that surround them. If some of the agnostics among us were to spend all their time with theists and were to marry theists and read books by smart theists, it is likely that they would change their minds. But this is enough to let us simply rephrase Pascal's argument: it is now an argument to start doing the things that would eventually lead to your having resolute belief in God. Since you can certainly do those things, it looks like once the argument is rephrased, there is no longer any relevant disanalogy between it and any ordinary decision problem that should be solved by computations of expected utility. We have indirect control over our beliefs, just as we do over other outcomes that are normatively relevant objects of decision-making.

Note also that this point allows us to avoid a different type of concern one might have had about Pascal's wager. One might have worried that Pascal is encouraging us to form belief in God for the wrong kind of reason: we shouldn't believe in God *just for gains to ourselves*, but rather because we think it's *actually true* that God exists. Arguably, when one goes about acquiring a belief in the more indirect way we're now envisaging, this *will happen* at least at a superficial level in consciousness: after hanging around long enough with smart theists, one will just see the case for God's existence differently, and this will be what most immediately causes one to have the belief, though an earlier choice did lead in an indirect way to this later state. This seems no different from causing oneself to collect new evidence, and then believing on the basis of that new evidence: one doesn't believe for the wrong reasons in this kind of case, and so, in our case, one doesn't either.

But there's a different problem with the Objection from Doxastic Involuntarism. The core idea behind this argument is that beliefs are unlike actions because beliefs aren't *voluntary*. It can't be because beliefs cannot be controlled at all, for we saw that indirect control is an option. But what was the reason for thinking that beliefs aren't voluntary? It was that we can't form beliefs at the snap of a finger. Beliefs aren't thumbs that we can put up or down. The underlying argument, then, seems to rest on the following presupposition:

> *The Voluntary => Just By Willing It Thesis*: One's A-ing is voluntary only if one can ensure whether one As or does otherwise just by willing it.

Alas, the Voluntary => Just By Willing It Thesis is false for at least a couple of reasons. The first is that there are lots of outcomes we can't ensure just by willing them but that we can still bring about voluntarily. Suppose that I want to finish writing a certain paper by 8:00, and decide to make this my goal. When I start off, there's no certainty about my succeeding: as I quite consciously realize, perhaps I'll get too tired or distracted or a disaster will happen. So, I can't *ensure* that I'll finish the paper by 8:00 *just by willing it to occur*. But, if I do succeed, I surely succeed voluntarily. If I finish, I finish through voluntary choice. So, the Voluntary => Just By Willing It Thesis must be rejected. But if it is, there is no longer clearly a disanalogy: I can also plan to acquire a certain belief by gathering evidence that will count in favor of its truth. Sure, I can't ensure it just by willing it. But that's not true of all voluntary acts anyway.

A different problem is that there are lots of negative outcomes that we voluntarily bring about even though we couldn't have ensured otherwise just by willing it. I cannot bring myself to stab myself in the stomach. No matter how hard I try, I just can't do it: if I bring the knife closer, I'll just eventually pull it away. In pulling the knife away and preventing myself from stabbing myself, do I act voluntarily? It seems like I do. But I couldn't have done otherwise just by willing it. I would have had to brainwash myself into radically suicidal thoughts to manage it. So, it's also false that one's A-ing is voluntary only if one can ensure whether one As or does otherwise just by willing it. This is related to the earlier point we saw in thinking about the virtuous beings God could have created. They could have stood in the same relation to their virtuous acts as I stood to my not stabbing myself: they also couldn't have done otherwise, but were still voluntary and free in their choices.

So, the upshot of all this is that the Objection from Doxastic Voluntarism fails, though it fails for interesting reasons that in some unseat some pre-reflectively attractive views about freedom, voluntariness, and states of mind like belief that we can't alter at the snap of a finger.

All the same, we shouldn't accept Pascal's wager, because there is a different and much more decisive objection to it which I'll call the *Cancellation by Overgeneralization Objection*.

As this name indicates, Pascal's wager is threatened by an embarrassing sort of overgeneralization. Consider the possibility of an Anti-God: a supernatural being who will send me to his version of heaven ("mirror heaven"), which is just as good as God's, if I disbelieve in God but believe in Anti-God, and who will send me to his version of hell ("mirror hell") if I believe in God but disbelieve in Anti-God. A precisely inverted decision matrix can be constructed to show that, as long as I assign a nonzero subjective probability to the prospect that Anti-God exists, it is more rational to believe in Anti-God and disbelieve in God than not to believe in Anti-God and believe in God:

| | **Anti-God exists** | **Anti-God doesn't exist** | **Expected Utility** |
|---|---|---|---|
| **Believe in God** | $(-\infty)(\text{s.pr.} > 0 < 1) = -\infty$ | $(x > 0)(\text{s.pr.} > 0 < 1) > 0 < \infty$ | $-\infty$ |
| **Don't believe in God** | $(+\infty)(\text{s.pr.} > 0 < 1) = +\infty$ | $(x > 0)(\text{s.pr.} > 0 < 1) > 0 < \infty$ | $+\infty$ |

If we put together this matrix with the previous one, the infinities cancel out. Consequently, we lack sufficient pragmatic reason to believe in God, since there is an (incompatible) alternative that would get us the same expected utility. By overgeneralizing, Pascal's argument undermines itself.

Now, it's certainly true that, for this objection to work, it has to be permissible for us to assign a nonzero subjective probability to the hypothesis that Anti-God exists. But, dialectically speaking, there is little of use that Pascal could say against this presupposition of the objection. Pascal starts off the presentation of his wager by *conceding* that there aren't any good epistemic reasons to believe that God exists. And on the face of it, there are pretty strong epistemic reasons to believe that he doesn't exist, since the Problem of Evil looks pretty much insoluble, and the argument for atheism from this problem was both deductively valid and apparently sound. If anything, the Anti-God hypothesis stands on better epistemic ground than Pascal thinks the God hypothesis does: since nothing is known about Anti-God except what he would do for us, we can't argue on empirical grounds that he clearly doesn't exist. We don't, of course, have positive epistemic reasons to believe that Anti-God exists, but, if Pascal is right, we also don't have positive epistemic reasons to believe that God does. So, the balance of epistemic reasons for the Anti-God hypothesis should actually more favorable by Pascal's own lights than the balance of epistemic reasons for the God hypothesis!

Accordingly, if *he* assumes – as he must – that we can assign a nonzero subjective probability to the hypothesis that God exists, he must *ipso facto* grant that we can assign a nonzero subjective probability to the hypothesis that Anti-God exists, since the latter hypothesis is on as good if not better epistemic ground than the former. Hence, the Cancellation by Overgeneralization Objection must look decisive to Pascal if he accepts the presuppositions of his own argument.

**MEETING 5**

## 1. Cartesian Epistemology, Part I: Principles, Knowledge and Doubt

A striking feature of the *Meditations* is that Descartes starts off by assuming without any argument some remarkably strong principles about when we are permitted to retain beliefs and when we ought to reject them. One might worry about how this fits with the surface-level attitude of the text, which is one of extreme doxastic humility: Descartes postures himself as being prepared to give up any of his beliefs if he thinks he has sufficient reason to do so, and to grant that he may very often have sufficient reason to do so. I'll return to this worry later.

For the moment, though, let's grant Descartes his principles. A crucial one that Descartes needs for much of his project is stated in the second paragraph of the First Meditation:

> My reason tells me that as well as withholding assent from propositions that are obviously false, I should also withhold it from ones that are not completely certain and indubitable. *So all I need, for the purpose of rejecting all my opinions, is to find in each of them at least <u>some</u> reason for doubt.* (1)

For the sake of briefly reconstructing some of Descartes' arguments, let's paraphrase the underlying principle in a crisper form:

> *The Indubitability Principle*: For any proposition P, one ought epistemically to withhold belief in P if one has a good epistemic reason for doubting that P.

This principle doesn't obviously have radical skeptical implications just on its own. After all, it is an open question just what counts as a good reason for doubting P. One could accept the Indubitability Principle but have an extremely demanding account of what it takes to be a good epistemic reason for doubting a claim. This may be the view that common sense pre-reflectively encourages in us: we really shouldn't believe that P if there is a sufficient epistemic reason for doubting P, but since there simply *aren't* such reasons for doubting much of what we believe, we don't end up dogged by radical skepticism.

But, at least at first, it looks like this can't be Descartes' tack. For until he relies on the existence of an undeceiving God to argue for the reliability of clear and distinct perception in the Third and Fourth Meditations, he takes himself to have good reasons for doubting all propositions about the external world. What *is* his reason for taking himself to have good epistemic reasons for doubting these propositions? If we reflect on the way he describes his skeptical scenarios – the "dreaming away life" scenario and the "evil demon" scenario – we find an answer.

A key thing he points out about these scenarios is that they are *experientially indistinguishable* from waking life as we typically suppose it to be. Since Descartes takes the indistinguishability of our experiences in waking life from our experiences in these skeptical scenarios to be a reason for doubt, he must be assuming the following principle about good epistemic reasons for doubt:

> *The Indistinguishability Principle*. If one would have indistinguishable experiences in two possible cases A and B, and these experiences are all one has to go on epistemically in deciding whether one is in A or in B, then one has a good epistemic reason to be doubtful (i.e., to suspend judgment) about which case one is in.

We can use the Indistinguishability Principle and the Indubitability Principle to fairly reconstruct the reasoning that leads Descartes to initial skeptical doubt. It goes like this:

1. An evil demon could have caused me to undergo all the experiences I've ever undergone, and could have deceived me into accepting the contents of these experiences even though they are all false and there is no external world at all.
2. If (1) is true, I couldn't distinguish the experiences I would have in the ordinary world where I typically suppose myself to be from the demon-induced ones.
3. All I have to go on in deciding whether I am in the world where I typically suppose myself to be are my experiences.
4. So, by the Indistinguishability Principle and (1 – 3), I have a good epistemic reason to be doubtful about whether I am in the demon-deception scenario or the world where I typically suppose myself to be.
5. And so, by the Indubitability Principle, I ought epistemically to withhold belief on the claim that I am in the world where I typically suppose myself to be. And that's just to say that I should be a skeptic about the external world.

One can actually simplify Descartes' reasoning so that it doesn't rely crucially on the Indubitability Principle. To do this, we could replace the Indistinguishability Principle from above with a stronger principle:

> *Revised Indistinguishability Principle.* If one would have indistinguishable experiences in two cases A and B, and these experiences are all one has to go on epistemically in deciding whether one is in A or B, then one has no epistemic reason to prefer the hypothesis that one is in A over the hypothesis that one is in B or *vice versa.*

A more contemporary version of the same skeptical argument would then look like this:

A. An evil demon could have caused me to undergo all the experiences I've ever undergone, and could have deceived me into accepting the contents of these experiences even though they are all false and there is no external world at all.
B. If one's experiences would be the same under one hypothesis H as under a different incompatible hypothesis H*, and H and H* could only be distinguished by experience, then, by the Revised Indistinguishability Principle, one has no epistemic reason to prefer H over H* or *vice versa.*
C. So, by (A) and (B), I have no more reason to believe the hypothesis that there is an external world that closely fits my experiences than I have to believe the hypothesis that I am being deceived by an evil demon, and no reason to prefer the former over the latter, since these hypotheses could only be decided by experience.
D. If I have no reason to prefer H to H* or *vice versa*, then I ought epistemically to suspend judgment on whether H or H* is true.
E. So, I ought epistemically to suspend judgment on whether the external world hypothesis or the demon hypothesis is true. And that's just to say that I should be an external world skeptic.

I think this is a slightly better argument than the one Descartes gives, because the Revised Indistinguishability Principle strikes me as being scarcely less plausible than the weaker principle it replaces, and it avoids reliance on the Indubitability Principle, which itself could be put into question. But given how crucial the Indubitability Principle is to Descartes, I don't think he'd accept my revision of his argument (though he really ought to!).

So, setting that quibble aside, where might either argument go wrong? Although the demon scenario may seem *odd*, it is surely *conceivable*. After all, in general, plenty of odd things are perfectly conceivable. If one really thinks one can't conceive of this scenario, or that its conceivability doesn't entail its possibility, we can just turn to different skeptical scenarios that are more clearly possible, like a variation on the "dreaming away life" scenario, or a case in which our brains are being stimulated by deranged neuroscientists into making us have precisely the experiences we are having now. Again, while such scenarios may strike us as odd, their *mere possibility* is all that is required for the argument to work. So, the first step of the Cartesian arguments we're considering probably isn't what's amiss.

What about premise (2) in the first argument? This premise is more plausibly challenged, and it's the premise that Descartes himself (perhaps surprisingly!) will have to reject. To bring out the challenge, let's note that there are two things one might mean by "distinguish". Suppose I'm looking at a Van Gogh painting and a molecule-for-molecule duplicate of the painting that I just created with some fancy futuristic machine. Upon taking the duplicate out of the machine, there is *one* sense in which I can't distinguish it from the real Van Gogh, because it *looks exactly the same*, and *another* sense in which I can, because I know that *I* just created it, not Van Gogh. Call the first sense "L-distinguishing", since just by *looking* I can't tell the difference, and the second sense "K-distinguishing", since by appealing to background *knowledge* I can tell the difference.

Now, it is clear that I cannot L-distinguish between the experiences I would have in the world in which I'm deceived by an evil demon and the experiences I would have in the world in which my experiences are completely accurate representations of reality. After all, by stipulation, they "look" the same. But we know in general that L-indistinguishability doesn't entail K-indistinguishability, and hence that the Indistinguishability Principle is false if 'indistinguishable' in it means 'L-indistinguishable'. For, after all, I *do* have a reason to prefer the hypothesis that the thing I just took out of the machine is not the real Van Gogh, though it is L-indistinguishable from the real Van Gogh. So, the friend of the skeptical argument must assume that 'distinguish' and 'indistinguishable' mean 'K-distinguish' and 'K-indistinguishable' in it.

But now it looks like the first reading of the argument *simply presupposes what it is trying to establish*: namely, that I cannot use the knowledge I have to decide whether I am in the real world or the demon world. The whole *point* of the skeptical argument was to show that I lack this knowledge, so it cannot simply *assume* that I lack it. So, the argument may seem to beg the question against the anti-skeptic, and hence to get nowhere. (As it happens, this is precisely the response to indistinguishability-based arguments for skepticism that people tend to defend these days. See Timothy Williamson's book *Knowledge and its Limits* for a major example.)

Still, a skeptic might fairly balk at this objection. His question for us will be *what knowledge we have on which to rely in K-distinguishing the demon world hypothesis from the ordinary world hypothesis*. He will ask us to *show him* how we could have this knowledge. How could we do that? It looks like the only thing we could do to respond to him would be to say that we know that sense perception is generally reliable. But how will we establish that result? We can't do it by appealing to science, since purported scientific knowledge presupposes the reliability of sense perception, and that would beg the question. It's hard to see how we could know it *a priori*: reflection alone can't tell us anything about how good our belief-forming faculties might be.

So, the objection we just considered to premise (2) may leave things just as they were, depending on how we understand the aims of the skeptic.

If it was the skeptic's goal to show that we lack knowledge *to our satisfaction*, then the burden of proof is on him to come up with some non-circular defense of premise (2). If, on the other hand, the burden of proof is *on us* to come up with a defense of our beliefs *to his satisfaction*, then we're going to be stuck with the very hard task of showing them to be non-circularly justifiable. As we'll see again in a moment, Descartes thinks he has a way to do this, but, given its reliance on theological premises that a fair number of us probably wouldn't accept, it's hard to see how it would help to show that all of us have the knowledge we pre-reflectively take ourselves to have. (Plus, as we'll also end up seeing, Descartes' route is itself arguably circular.)

Before turning to that, let's consider some other options for dealing with these skeptical arguments. What about (3) and the corresponding assumption built into (C) that we could only decide between the demon hypothesis and the external world hypothesis *by experience*?

This premise is surely questionable, though it is unclear how much we can achieve by questioning it. To see the reasons for doubt, consider the fact that scientists are often stuck with multiple hypotheses that equally well predict the observational data. What do scientists do when they're stuck in this way? Well, besides trying to figure out ways to gather more observational data, one thing they do is to consider the *comparative simplicity* of the theories and their *explanatory quality*, and tentatively stick with the simpler and explanatorily better theory. We normally think that they can be intellectually responsible in doing this. This might be wrong, but if it were right, one might think one could appeal to similar considerations in deciding between various skeptical hypotheses and the hypothesis that things are for the most part as they appear.

The main issue with this proposal is that it's hard to see how it could handle all skeptical hypotheses. Some versions of the demon hypothesis actually seem *simpler* than the hypothesis that things are for the most part as they appear to be. If the demon is just a purely mental being and I am too (as Descartes thinks, since he thinks, by appeal to the analogy with the ball of wax, that our physical properties are inessential to our being), and we're just floating in some spirituous realm, then all that would exist would be a couple of mental things and the experiences they have. If, on the other hand, things are for the most part as they appear to be, things will be *a lot more complicated*: there will be billions upon billions of simple physical things and lots and lots of mental things too. So, the normal world hypothesis looks vastly more complex, at least on one sensible measure of complexity, than the demon hypothesis. Barring an argument for some different measure of complexity or for the impossibility of this more extreme version of the demon case, it looks like this attempt to reject (3) is hopeless.

This leaves us with little else to attack in the Cartesian skeptical arguments. One last possible place to put pressure is on the core assumption of both arguments that *whether one is epistemically permitted to believe something depends exclusively on the epistemic reasons one has for believing it*. This idea is clearly presupposed by the Indubitability Principle, but it is also presupposed by the weaker assumption (D) in the second argument. How plausible this idea is depends to some extent on what one takes an epistemic reason for belief to be. Suppose one accepts:

> *The Argumentative Theory of Epistemic Reasons for Believing.* One has an epistemic reason to believe that P only if one has a sound deductive argument or a good inductive or abductive argument for believing that P.

If we accepted this theory of epistemic reasons for belief, then the idea that epistemically permissible belief requires reasons will not be plausible. I certainly think I'm epistemically

permitted to believe elementary algebraic truths like "a + b = b + a". I don't have to present deductive, inductive or abductive *arguments* for these claims, because they are just *obviously* true.

Of course, one might reject this account of the nature of epistemic reasons for belief. One could say that the obviousness of some belief is itself a good reason for that belief, quite independently of any arguments. But once one does this, the skeptical argument may seem to have less force, at least depending on how we understand the broader dialectic in which it is embedded. A stubborn defender of common sense might just say: "It's as obvious to me that I have hands as it is that 2 + 2 = 4. If obviousness can be a reason to believe the latter, it can also be a reason to believe the former. And that's enough to show that the skeptical argument must fail." This was the strategy of the early 20th century philosopher G. E. Moore: he thought that the premises of any skeptical argument could never be more plausible than the claim that he has hands, and so there's just no reason to let the skeptic tempt us with doubt in the first place.

Obviously, this isn't going to be satisfactory if the task was to show the skeptic *on his own terms* that we have knowledge. But perhaps – and this is what philosophers like Moore insist – we should refuse to meet the skeptic on his own terms. It's the *skeptic's* job to show *us* that we're mistaken, and he's got to do that by coming up with an argument that rests on premises that are more obvious to us than the denial of the skeptical argument.

This is another response to skepticism that many people today find attractive. But it may not help those of us who were slightly compelled by Descartes' worries in the first place. (Of course, the claim that we *should* be compelled by his worries may require some of the assumptions that this response makes: after all, Descartes doesn't give us any argument for the Indubitability Principle or the (Revised) Indistinguishability Principle: he just *presupposes* them, presumably because he thinks they're obvious!) So perhaps we should then turn to see what Descartes himself has to say about the skeptical arguments from the First Meditation. As we'll see, it's unclear that his response is really more compelling than Moore's or Williamson's.

## 2. Cartesian Epistemology, Part II: Clarity, Distinctness and the Idea of God

The Second Meditation famously does not furnish Descartes with much of a foundation for the rest of our knowledge. He discovers one claim that is indubitable – namely, that he exists. How does he propose to expand the foundations to avoid radical skepticism? At first, it looks like his strategy in the Third Meditation is to isolate the property that explained this claim's indubitability, and to then argue that there are enough further claims that have this property to get us the rest of our knowledge. What is the property? Here is his answer:

> Now I will look more carefully to see whether I have overlooked other facts about myself. I am certain that I am a thinking thing. Doesn't that tell me what it takes for me to be certain about anything? In this first item of knowledge there is simply a vivid and clear perception of what I am asserting; this wouldn't be enough to make me certain of its truth if it could ever turn out that something that I perceived so vividly and clearly was false. So I now seem to be able to lay it down as a general rule that whatever I perceive very vividly and clearly is true. (9)

This is a puzzling passage, given what Descartes goes on to say. A few paragraphs later, he says:

> But what about when I was considering something…in arithmetic or geometry, for example that two plus three makes five? Didn't I see these things clearly enough to accept them as true? Indeed, the only reason I could find for doubting them was

> this: perhaps some God could have made me so as to be deceived even in those matters that seemed most obvious. Whenever I bring to mind my old belief in the supreme power of God, I have to admit that God could, if he wanted to, easily make me go wrong even about things that I think I see perfectly clearly. (10)

The issue here is this. Perhaps Descartes is right that clear and distinct perceptions are always true. But how does he know *whether he's having a clear and distinct perception*? It looks like he's willing to admit that he is sometimes wrong when he believes that he clearly and distinctly perceives something. So how does he know he isn't wrong in the cases where he uses this rule?

At first, it seems like this worry doesn't really matter. Descartes goes on to say that he has an idea of God, and that this idea represents God as being infinite. We could grant to Descartes that this claim – i.e., that he has this idea – is on as firm footing as the *cogito* (i.e., the claim that he thinks and so exists). For it does seem impossible to see how one could be wrong about *having ideas*: in order to be deceived into thinking one has an idea, one has to have some idea, just like in order to be deceived, one has to exist. It looks like we can set aside the business about clear and distinct perception and note that the claim that Descartes has an idea of God has a *different* property that the *cogito* had: namely, the property that *if one tries to doubt this claim, the claim has to be true. This* is what seemed good about the *cogito*, and it looks like we can understand this property without appealing to clear and distinct perception. Perhaps Descartes made a mistake in insisting that what was special about the *cogito* was that it was clearly and distinctly perceived.

The problem, though, is that something like clear and distinct perception seems to be required for the rest of the argument to work. The argument doesn't appeal *just* to the assumption that we have an idea of God. It also appeals to a *principle*, which Descartes summarizes as follows:

> Now it is obvious by the natural light that the total cause of something must contain at least as much reality as does the effect. For where could the effect get its reality from if not from the case? And how could the cause give reality to the effect unless it first had that reality itself? Two things follow from this: that something can't arise from nothing, and that what is more perfect – that is, contains in itself more reality – can't arise from what is less perfect. (12)

One issue is that the principle to which Descartes adverts is not indubitable, and is not at all like the *cogito*: it is *not* true that if you try to doubt this principle, it must be true. Indeed, the principle seems either false or uninteresting, depending on how we understand it. Notice that there is a distinction between two claims: (i) the idea of God is about something that is infinite, (ii) the idea of God itself is itself infinite. If (i) is the claim that Descartes accepts, then his principle either does not establish the conclusion or must be false: it is *not* true that if an idea is *about something that has property F to degree D*, then the cause of that idea *must itself have F to a degree equal to or greater than D*. I have an idea of a long bridge. But it hardly follows that the idea is itself at least as long as the bridge! So, either the principle is false or (i) isn't what Descartes intends to appeal to. But it's hard to see how he can appeal to (ii), since my idea of God, though it is *about something infinite*, is not *itself infinite*: after all, my mind contains this idea, and my mind is finite. So, Descartes must have intended (i) rather than (ii), and if he did, we can cast doubt on the correspondingly required reading of his principle. So, to sum this up, if the principle is to do any work in Descartes' argument, it is going to end up being a dubious principle whose truth is not established by our doubting it, and hence will not be like the *cogito*.

But there's a different problem in Descartes' reasoning. Notice how Descartes supports his principle: he says that it is made "obvious by the natural light". What is the natural light? Whatever it is, it can't be something which, when cast on some claim, renders that claim indubitable, since Descartes' principle is dubitable. Of course, it *could* be something that is simply *reliable* though not always such that its outputs are indubitable. In this way, it could be like clear and distinct perception. But how can Descartes permissibly rely on this mysterious faculty of "natural light"? If he relies on it without establishing its reliability, then it's incredibly hard to see why we should have *ever* taken his earlier arguments seriously. After all, our faculty of judgment might itself turn out to be reliable (Descartes after all thinks it is), and if all that matters for our being permitted to believe its outputs is its simply *being* reliable, then there is no longer any point in even trying to take skepticism seriously: we could have good enough reason to accept the verdicts of our faculty of judgment just by relying on *it*, because it's *in fact* reliable.

If, on the other hand, Descartes can only permissibly rely on this faculty by *first* establishing that it is reliable, then he's in a lot of trouble, because it looks like the only way he could establish that is by, well, *relying on clear and distinct perception*. But the *point* of his appeal to the concept of God, and to the principle about the nature of causation, is to establish that clear and distinct perception *is* generally reliable. After all, in the quote from p.10 that we considered earlier, Descartes seems to concede that until he's established that there is a well-meaning God, he cannot trust himself when he thinks he is clearly and distinctly perceiving something, because he could imagine feeling like he is doing so when he really is not. Without establishing that, then, he can't permissibly rely on clear and distinct perception.

There are related and even simpler problems. One of them is that it looks clearly circular for Descartes to assert the rule that clear and distinct perception is generally reliable only to go on trying to prove this rule by relying on God's existence. (This is the famous "Cartesian circle", which is much discussed in the secondary literature.) Another is that it's completely unclear what the difference between "natural light" and "clear and distinct perception" is supposed to be. Unless there is a difference, Descartes' argument looks circular in a further, even worse way.

In any case, the broader and more worrying upshot of all this is just that for the argument in the Third and Fourth Meditations to work, Descartes is either going to have to make a bedrock appeal to something that just seems obvious *even though it's dubitable* and *implicitly* take it for granted that this is good enough for knowledge, or *explicitly* concede that some of our bedrock knowledge is not actually indubitable, which would conflict with the Indubitability Principle and the assumptions Descartes makes about the connection between doubt and knowledge.

Perhaps he will just accept the principle about causes, or the principle that clear and distinct perception is generally reliable, or the claim that the "natural light" is a distinct faculty from clear and distinct perception that is generally reliable. If he goes for any of these options, and yet admits that he might sometimes be wrong about whether he really *is* relying on clear and distinct perception or on the natural light, then he has to grant either (i) that it is permissible to rely on a rule *simply because it's reliable* when one can in fact doubt whether one is applying the rule, or (ii) that we can be permitted to believe some things that are not indubitable. Admitting (ii) would undermine the whole project, since it relies on the Indubitability Principle. Admitting (i) would call into doubt the point of the project, because as long as our faculty of judgment is itself reliable, that ought to be enough, given (i), to permit one to rely on it.

**MEETING 6**

## 1. Locke's Epistemology and Philosophy of Mind

While there's a multitude of views in Locke's *Essay* worth discussing, we've been focusing on the following four in his epistemology and philosophy of mind:

*Epistemological Views*

I. Fallibilism. Locke's epistemology departs significantly from Descartes' in large part because Locke recognizes the importance of epistemically rational but *fallible* (i.e., possibly mistaken) belief. For this reason I'll call him a *fallibilist*. Recall that Descartes presupposed without argument that we ought epistemically to suspend judgment on any proposition that isn't completely certain and indubitable, or that isn't nontrivially demonstrable from propositions that are completely certain and indubitable. Locke thinks this presupposition is much too extreme and should be rejected. He thinks we permissibly hold many beliefs that fall short of knowledge.

Still, he in a different way agrees with Descartes: he grants that anything properly called 'knowledge' requires certainty. This is obvious in passages like:

> *Our Knowledge being short, we want something else.* The understanding faculties being given to man, not barely for speculation, but also for the conduct of his life, man would be at a great loss if he had nothing to direct him but what has *the certainty of true knowledge*. For that being very short and scanty, as we have seen, he would be often utterly in the dark, and in most actions of his life, perfectly at a stand, had he nothing to guide him in the absence of clear and certain knowledge…. (Book V, chapter xiv, section 1)

> *[Probability] is to supply the Want of Knowledge.* Our knowledge, as has been shown, being *very narrow*, and we not happy enough to find certain truth in everything which we have occasion to consider, most of the propositions we think, reason, discourse, nay, act upon, are such as we cannot have undoubted knowledge of their truth; yet some border so near upon certainty that we make no doubt at all about them, but assent to them as firmly, and act, according to that assent, as resolutely as if they were infallibly demonstrated, and that our knowledge of them was perfect and certain. But there being degrees herein, from the very neighborhood of certainty and demonstration, quite down to improbability and unlikeliness, even to the confines of impossibility, and also degrees of assent from full assurance, quite down to conjecture, doubt, and distrust.... (Book V, chapter xv, section 1)

But while Locke seems to agree with Descartes that knowledge requires certainty, he also thinks that we are permissibly certain of *a lot more* than Descartes. This is clear in passages like:

> If we persuade ourselves that our faculties act and inform us right concerning the existence of those objects that affect them, it cannot pass for an ill-grounded confidence; for I think nobody can, in earnest, be so skeptical as to be uncertain of the existence of those things which he sees and feels…. As to myself, I think God has given me assurance enough of the existence of things without me; since, by their different application, I can produce in myself both pleasure and pain, which is one great concernment of my present state. This is certain: the confidence that our faculties do not here deceive us is the greatest assurance we are capable of concerning the existence of material beings. (Book V, chapter xi, section 3)

Descartes would have only granted the title of certainty to propositions like "I think and therefore exist" and "I have an idea of God", ones that are entirely based on private, internal states. In this passage, Locke is suggesting that it would be foolish to deny that we can *properly* have nearly the same level of certainty in publically observable propositions about the external world like "There is a piece of paper on this table".

II. Externalism. Another respect in which Locke departs significantly from Descartes is in denying that we have to have *independent validation* of the reliability of a belief-forming process to be able to justifiedly believe the deliverances of that process, and even to have knowledge by way of it. In this respect Locke is what contemporary philosophers would call an *externalist* about knowledge and justified belief. Externalists characteristically deny claims like the following:

> *KK Thesis*. In order to have knowledge (or justified belief) that P *via* some belief-forming method M, one must have independent knowledge (or justified belief) of the reliability of M. *One must know (by independent means) that one knows in order to know.*
>
> *No Method-Circular Response Thesis*. In order to undermine doubt about some belief-forming method M, it is not permissible simply to rely on M.

That Locke rejects these theses, and hence endorses a kind of externalism, is clear for at least two reasons. One is that his response to skeptical worries about perceptual knowledge relies explicitly on perceptual knowledge. To see this, recall the first thing he says in replying to the skeptical worry that our ideas do not come from an external world:

> [I]t is plain those perceptions are produced in us by exterior causes affecting our senses; because those that want the *organs* of any sense never can have the ideas belonging to that sense produced in their minds. This is too evident to be doubted; and therefore we cannot but be assured that they come in by the organs of that sense, and no other way. The organs themselves, it is plain, do not produce them; for then the eyes of a man in the dark would produce colors, and his nose smell roses in the winter; but we <u>see</u> nobody gets the relish of a pineapple, till he goes to the Indies, where it is, and tastes it. (Book V, chapter xi, section 4)

This reply relies on perceptual knowledge. Locke is telling us that the reason why it's obvious that our ideas come from an external world is that, well, we can <u>*see*</u> that they do. (This isn't the *only* reason he gives: he also thinks it's just impossible that the mind could generate these ideas by itself. But that very claim is an *empirical* claim for Locke, not a conceptual claim.)

The only way Locke could say things like this without being incoherent (or just plain silly) is by implicitly rejecting the KK Thesis and the No Method-Circular Response Thesis. Locke's reply to the skeptic is a lot like G. E. Moore's, which I mentioned in passing last week in discussing objections to Descartes: he thinks it is *already* "too evident to be doubted" that our ideas are produced by external objects *for reasons that essentially presuppose perceptual knowledge*.

Of course, you might complain that Locke is simply not responding to the skeptic at all. But it's worth recalling something from last time to see that this would be unfair – i.e., that there are two very different ways of understanding the dialectic between the epistemologist and the skeptic.
The first is *modest*. The skeptic could be trying to convince the common sense epistemologist, using premises *the latter* would be willing to accept, that he lacks knowledge and justified beliefs about the external world. As philosophers like G. E. Moore point out, the skeptic is bound to

lose *this* dialectical game, because we will *never* be as certain of the skepticism-generating premises as we are of the denial of the skeptical argument's conclusion: it's *more obvious* to us that we know we have hands than that, say, rational belief requires certainty, or that the KK Thesis is true.

The second construal of the dialectic is, to put it mildly, less modest. It might be thought that *we* have to show the *skeptic*, on premises that *he* would accept, that we have knowledge and justified beliefs about the external world. The skeptic is bound to win *this* dialectical game, because the fact is that we're going to have to stop somewhere in replying to him, and in stopping, we're going to have to stop by appealing to a method for which we have no *independent* warrant.

This really does seem inevitable for simple reasons – a fact that shows, I think, that the less modest construal of the dialectic is in fact *immodest*. Every system of deductive logic has *basic inference rules*: unproven rules that are used to prove everything else. (A typical example is *modus ponens*: from P, and P → Q, infer Q. This is a basic rule of most classical logics.) One can, of course, try to engage in metalogical proofs that show that the basic inference rules can prove all and only logical truths. But these metalogical proofs inevitably use the basic rules! I've never seen soundness and completeness proofs for a classical system that didn't rely on *modus ponens*. So, insofar as a *paradigm* case of absolutely certain knowledge – i.e., deductive proofs of logical truths – is a case where we can't have independent validation of the reliability of the belief-forming rules that are used, I think we should conclude that there's no point in trying to argue with the skeptic on "less modest" grounds. That's just asking too much. We don't need that kind of validation to be able to properly use *modus ponens*. And so we ought, just like Locke, to reject the KK Thesis and the No Method-Circular Response Thesis.

Locke is charitably read as engaging in a modest reply to the skeptic, and reasonably so. He's suggesting that the skeptic will never be able to convince us using premises that *we* would be willing to accept that we lack knowledge and justified belief about some propositions about the external world. And in doing so, he is simply trusting in the reliability of our faculties, and *ipso facto* implicitly rejecting the KK Thesis and the No Method-Circular Response Thesis.

In fact, he at times seems to be quite *explicit* about this, and this is my second reason for reading Locke as an externalist. Consider:

> [W[e cannot act anything but by our faculties; nor talk of knowledge itself, <u>but by the help of those faculties which are fitted to apprehend even what knowledge is</u>. But <u>besides the assurance we have from our senses themselves</u>, that they do not err in the information they give us of the existence of things without us, when they are affected by them, we are further confirmed in this assurance by other reasons…. (Book V, chapter xi, section 3)

Here Locke explicitly says that we may (and must!) rely on our faculties in order to appreciate their reliability, and that "our senses themselves" can provide this "assurance". This is a paradigmatically externalist claim, and a rejection of the No Method-Circular Response Thesis.

*Views in the Philosophy of Mind*

So much for Locke's epistemological views. Let's turn to the aspect of his system on which James focused a bit more in the lecture: his philosophy of mind.

III. <u>Concept Empiricism</u>. This part of Locke's philosophy of mind marks an even more conspicuous and self-conscious departure from Descartes and other rationalist thinkers in the

17th century (Leibniz, Spinoza and Malebranche). Locke thinks that all our simple concepts (or, in his language, "ideas") are derived from experience, either via *sensation* or *reflection*. So, he denies that there are innate concepts, which is a key tenet of rationalism and a key target of empiricism. This is clear throughout the text, but particularly so in this famous passage:

> Let us then suppose the mind to be, as we say, white paper, void of all characters, without any ideas; how comes it to be furnished? Whence comes it by that vast store which the busy and boundless fancy of man has painted on it with an almost endless variety? Whence has it all the materials of reason and knowledge? To this I answer, in one word, from *experience*. In that all our knowledge is founded, and from that it ultimately derives itself. (Book II, chapter i, section 2)

Here Locke actually goes beyond asserting just *concept empiricism*. He asserts a further claim that we can call *knowledge empiricism*, according to which all our knowledge is rationally derived from experience *in the sense that we need to consult experience in order to have reasons to believe what we believe*. Knowledge empiricism so defined doesn't actually follow straightforwardly from concept empiricism, and Locke is a little sloppy to make the inference so quickly.

Why doesn't it? Well, it's not crazy to claim, as the concept empiricist does, that we couldn't entertain any propositions without the help of experience. But it's not valid to infer from this that we must *consult* experience in order to *verify* every proposition: to say that knowledge requires us to *have* some experiences is not the same as saying that knowledge is *rationally grounded in experience*. To see the difference between these thoughts, consider the following claims:

A. There is a table here.
B. Either there is a table here or there is not a table here.

I have to *consult* experience to know (A): my reason for believing (A) is that I *see* that there's a table. But I can know (B) at some time *t* without *consulting* experience at *t*. I don't have to *check* whether it's true by *checking* whether one of the two disjuncts is true. For I know that (B) is an instance of an axiom of deductive logic that is obviously true *in virtue of its form*: namely, the law of excluded middle, "$P \vee \neg P$". Still, I couldn't have even entertained the proposition expressed by (B) without gaining the concept of a table through experience. So, from the claim that we couldn't entertain a proposition except *via* some antecedent experience, it doesn't follow that our knowledge of this proposition is *rationally grounded in experience*. My *reason* for believing (B) isn't an experiential reason like my reason for believing (A) is: it's instead that (B) is clearly an instance of a propositional schema that is guaranteed to be true in virtue of its *form*.

So, we should reject Locke's inference from concept empiricism to knowledge empiricism, and we should probably reject knowledge empiricism too, because my knowledge of logic isn't provided by *experiential reasons*, but instead by reasons generated by reliable *rational intuition*. Even so, should we still follow Locke in endorsing concept empiricism?

I think we shouldn't, though the reason why we shouldn't doesn't exactly spell a *huge* triumph for rationalism. Here is a simple argument for thinking that the concept of *extension* has to be innate:

1. In order to see something *as having a property F*, I have to antecedently possess the concept of F-ness.
2. There was never a time when, upon first viewing some external object, I failed to see it as having extension.

3. So, there was never a time when I failed to antecedently possess the concept of extension.
4. If (3), my possession of concept of extension predates my having any perceptual experiences at all (e.g., seeing something as extended).
5. If my possession of a concept predates my having any experiences at all, then concept empiricism is false.
6. So, concept empiricism is false.

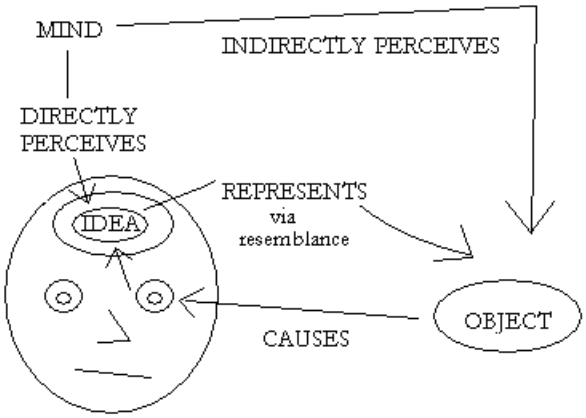
Notice that this argument doesn't overgeneralize to show that *all* concepts are innate. When I first saw a dog, I didn't recognize it *as a dog*: I had to learn that concept by being told about the nature of dogs, and to learn to reliably correlate this information with certain patterns of visual experience. The concept *dog* is thus not *built into* experience in the way that the concept of extension is. Indeed, most of my concepts are like this: I had to acquire some information and learn how to reliably correlate that information with certain patterns of experience in order to fully possess those concepts. By the lights of the argument just given, all these concepts will be derived from experience, together with some antecedent abilities, like the ability to quickly correlate certain information with certain patterns of experience.

So it's not as if a really strong kind of rationalism like Plato's will be true. (Plato thought our souls were in contact with all the basic Forms before we became embodied. When we became embodied, we lost the ability to remember these Forms. But all experience does is "remind" us about them: it doesn't help us to *acquire* the concepts of these Forms, but to simply help us dig them out of the "corporeally occluded" basements of our souls.) But concept empiricism of the very strong kind Locke does hold will indeed be false: his claim was supposed to be the general one that *all* concepts are acquired by experience.

The point I've just made is essentially the one that Immanuel Kant made later in the 18th century. Kant's view – a plausible view, I think – was that there are certain highly general, basic concepts that are *preconditions for our having any experiences at all*. We can't have visual experiences without automatically seeing things as being extended, and so possession of the concept of *extension* is a precondition for experience. Experience is, in other words, *conceptually structured*, and the concepts that structure experience (which Kant called "the Categories") are all innate.

This type of view is now pretty widely held among cognitive scientists, and has indeed been experimentally confirmed. (More surprising versions of it have also been experimentally confirmed. A theme of linguistics after Chomsky is a kind of "poverty of the stimulus" argument for nativism about grammatical concepts: basic syntactic concepts are innate, since there's no other way to explain how we could acquire languages with so little stimulus.) To this extent we're probably going to have to give up on a core tenet of Locke's view and become scientifically informed *moderate rationalists* like Kant.

IV. Indirect (Causal) Theory of Perception and the Primary/Secondary Distinction. The last major element of Locke's system as we've seen it is what contemporary philosophers would call an *indirect theory of perception*. Locke thinks that the only things we *directly* perceive are our ideas. We *indirectly* perceive objects in virtue of their causally contributing to our formation of these ideas, some of which *resemble* qualities in the objects. In a silly but perhaps useful illustration:




So, again, you *directly* perceive your ideas, which are caused by a combination of external factors (e.g., light) and internal factors (perceptual processing mechanisms), and they represent those objects *via* resemblance. In virtue of all this, the objects are indirectly perceived.

Locke adds an important qualification to this picture. Not all ideas resemble the qualities in objects that cause them. Only ideas of *primary qualities* do. Primary qualities are defined somewhat messily by Locke as follows:

> First, such as are utterly inseparable from the body, in what state soever it be; and such as in all the alterations and changes it suffers, all the force can be used upon it, it constantly keeps; and such as sense constantly finds in every particle of matter which has bulk enough to be perceived; and the mind finds inseparable from every particle of matter, though less than to make itself be singly perceived by our senses: e.g., take a grain of wheat, divide it into two parts: each part still has *solidity*, *extension*, *figure* and *mobility*; divide it again, and it retains the same qualities; and so divide it on, till the parts become insensible; they must retain still each of them all those qualities. For division…can never take away either solidity, extension, figure, or mobility from any body…. These I call *original* or *primary* qualities of body…. (Book II, chapter viii, section 9)

These are contrasted with *secondary qualities*, which are "nothing in the objects themselves but powers to produce various sensations in us by their primary qualities…." (II.viii.10) It's rather unclear in the text just why Locke thinks our ideas of primary qualities resemble them.

He doesn't really give an argument for this. He does give a series of related arguments to the effect that our ideas of secondary qualities (e.g., color, smell, sound) do *not* resemble anything that actually inheres in the objects that cause these ideas. Here is his example for color:

> Let us consider the red and white colors in porphyry. Hinder light but from striking on it, and its colors vanish; it no longer produces any such ideas in us; upon the return of light it produces these appearances on us again. Can anyone think any real alterations are made in the porphyry by the presence or absence of light, and that those ideas of whiteness and redness are really in porphyry in the light, when it is plain *it has no color in the dark*? It has, indeed, such a configuration of particles, both night and day, as are apt…to produce in us the idea of redness, and from others the idea of whiteness; but whiteness and redness are not in it at any time, but such a texture that hath the power to produce such a sensation in us. (II.viii.19)

The underlying argument here is usefully summarized as follows:

*Argument from Perceptual Relativity*

1. Depending on very particular circumstances of perception, the very same object will appear to have very different secondary qualities. (For example, by varying the lighting conditions, one can make an object appear to have very different colors.)
2. But differences in the circumstances of perception cannot change the qualities that an object has in itself; the *object itself* does not change as the circumstances of perception change.
3. So, secondary qualities like color are not qualities that the object itself possesses. (All it really possesses are powers to produce different ideas relative to different perceptual circumstances.)

A serious question for Locke is whether this type of argument overgeneralizes to show that *none* of our ideas actually represent external objects.

## 2. Berkeley's Idealism

This is precisely the question that leads Berkeley to suggest that we simply abandon the distinction between primary and secondary qualities. For, as he objects to an analogue of Locke ("Hylas") by playing the role of Philonous in his dialogues, what Berkeley in effect notes is that *every* quality can be subjected to the Argument from Perceptual Relativity.

We can alter our perceptual circumstances or our perceptual systems so that the same object will produce in us very different ideas of shape, size, mass, extension, and so on. Some of these variations are obvious ones: we can move around a table or look at it from different angles or distances, and it will appear to have a different shape and size. Some of the variations would have to be more radical: by taking various drugs, we could make it the case that, when we encounter some object, it produces in us a very different sensation of mass or even extension. Maybe the changes don't even need to be that radical to get Arguments from Perceptual Relativity to run in these cases. If you spend all day lifting huge boulders, chances are that various things that used to feel very heavy to you now feel significantly lighter.

What does Berkeley conclude from all this? He concludes that there really are no mind-independent qualities: the qualities that exist in the world are all just "possibilities of sensation" (to use J. S. Mill's nice phrase). This is a view called *idealism*. (This sense of "idealism" is technical and stipulative, and is *not* associated with the ordinary sense the word has when we talk about "idealistic people": in our context "idealism" is just the claim that the only things that exist are ideas.) And if what the original Argument from Perceptual Relativity was supposed to show was that a certain quality did not in fact reside in any external object, this should come as no surprise. *All* qualities are secondary qualities, if Berkeley's right. If secondary qualities really don't inhere in objects, then they're all "in us" rather than in some external world.

But here I think Berkeley is being too clever for his own good. He is taking advantage of Locke's own sloppiness in stating his view. There are two different ways of reading Locke. On one, secondary qualities *are* qualities of external objects: they just aren't the qualities *we typically take them to be*. They are *really* just powers to produce various sensations in us. The objects *do* have these powers in themselves: after all, when we change the lighting from color X to color Y, the object still remains such that *if* we were to switch the lighting back to color X, it would produce in us the same idea that it had been producing before. What the object *doesn't* mind-

independently have is anything that *resembles* the color *sensation*. But it still does have a property: the property of causing this kind of sensation in certain circumstances. On the second (crazy) reading, Locke thinks that by showing that secondary qualities are subject to perceptual relativity, he is showing that they are not properties of objects *at all*, not even *different* properties, such as the causal power to produce certain sensations. Locke sometimes does talk as though this is what he has in mind. But I think this is just sloppiness: he really *does* think there are genuine properties corresponding to secondary qualities. It's just that these qualities are quite different from what we naively think they are: they don't really resemble our ideas at all.

If we insist on the first reading of Locke, Berkeley's first argument for idealism fails. What it shows is just that a *different* assumption that Locke made is false: namely, the assumption that our ideas of primary qualities *resemble* these qualities, and that the *difference* between primary and secondary qualities is that the latter are powers to produce certain ideas whereas the former are not. What Berkeley *really* shows (if he shows anything) by pointing out that the Argument from Perceptual Relativity generalizes is that primary qualities are *also* just powers of external objects to produce certain sensations in us. He doesn't successfully show that there are no external objects at all: just that *all* of our ideas of them *fail to resemble them*. (This, by the way, is close to Immanuel Kant's view: there is a mind-independent world, but we can't know anything about it, because we could never know whether any of our ideas actually resemble it.)

Of course, Berkeley has a second argument for idealism. But it's a bad argument. Here it is:

1. We cannot conceive of anything and not conceive of it as being perceived (by someone or other).
2. Inconceivability entails impossibility.
3. So, it is impossible for anything to exist unperceived.

There is a dilemma for this argument.

"Inconceivable" might mean two different things. One way I can conceive of something is to engage in perceptual imagination: I can imagine what it would be like to interact with that thing. Accordingly, "inconceivable" might mean "impossible to entertain *via* perceptual imagination". Unfortunately, there are lots of things that are definitely possible that cannot be entertained by perceptual imagination. I cannot entertain by perceptual imagination what it would be like for the Universe to be infinitely temporally extended in a forwards direction: all of my imaginings will take place in a finite amount of time, and so they could never cover the "whole" span. But this is surely possible. So, if "inconceivable" means "impossible to entertain *via* perceptual imagination", premise (2) is false. "Inconceivable" could mean something different. Sometimes I conceive of things just by supposing for the sake of argument that they are true, and seeing whether a contradiction follows. If no contradiction follows, my supposition is logically consistent, and hence logically possible. "State of affairs X is inconceivable" could thus mean "supposing that X obtains entails a contradiction". But if *that's* what "inconceivable means", then premise (1) in the argument is false, because it *is* possible to entertain the thought that something exists unperceived without landing yourself in a contradiction.

So, either (1) or (2) is false, depending on the meaning of "inconceivable".

**MEETING 7**

## 1. The Traditional Problem of Induction

One of the most fascinating and vexing problems we've inherited from Hume takes the form of the following argument to the effect that induction cannot be justified:

1. To show that any particular use of induction is justified, we would need to show that, in general, it is rational to infer an instance of

   *The Projection Schema*: All unobserved Fs are Gs.

   from an instance of

   *The Sample Schema*: All observed Fs are Gs.

2. To show that it is rational to infer an instance of the *Projection Schema* from an instance of the *Sample Schema*, we would have to establish a *Uniformity Principle* according to which observed patterns extend into the unobserved world.
3. The argument for this Uniformity Principle could take two forms: it could be an *a priori* argument, or an *a posteriori* argument.
4. There could not be a good *a priori* argument for the rationality of inferring an instance of the *Projection Schema* from an instance of the *Sample Schema*. An *a priori* argument would show that the former is a *necessary consequence* of the latter. This can't be done: it is *logically possible* for the former to be false while the latter is true, and so it cannot be *a priori* guaranteed that the former is true when the latter is true. And if this logical relation cannot be *a priori* guaranteed, it follows that there is no *a priori* argument for the rationality of the inference.
5. There also could not be a good *a posteriori* argument for the rationality of inferring an instance of the *Projection Schema* from an instance of the *Sample Schema*. After all, an *a posteriori* argument will be an inductive argument. But we cannot use induction to justify induction: this is circular, and a circular justification is no justification at all.
6. So, since there could be no good argument for the rationality of inferring an instance of the *Projection Schema* from an instance of the *Sample Schema*, we cannot show that, in general, it is rational to make this kind of inference.
7. So, we also cannot show that any particular use of induction is justified.

On traditional interpretations, Hume went farther than just (7). He took (7) to establish:

8. None of our particular inductively formed beliefs is justified.

Notice that the inference from (7) to (8) is extremely nontrivial. It presupposes a principle that we saw in Descartes, and that Locke certainly seems to have rejected. The principle is:

> *The JJ Thesis*: In order to be justified in believing that P *via* some method M, one must have a good noncircular argument for believing that M is a reliable method.

The JJ Thesis leads not just to inductive skepticism, but also to external world skepticism. For it implies that in order to be justified in believing that P *via* perception, we must have a good noncircular argument for believing that perception is a reliable belief-forming process. That

argument isn't easy to come by: Descartes tries to pull it off, but even he lands in circularity. Without such an argument, the JJ Thesis implies that none of our perceptual beliefs is justified. A sensible conclusion to draw from this is simply that the JJ Thesis is too strong, and should itself be rejected. It sets a standard that is impossible to meet. In general, we think that 'ought' implies 'can': if we ought to live up to some standard, we can live up to it. If the JJ Thesis is supposed to express a standard that we ought to live up to in order to have justified beliefs, then typical 'ought' implies 'can' reasoning should lead us to reject it.

The JJ Thesis also implies skepticism about deductively formed beliefs. For it implies:

> *JJ-Modus Ponens*. In order to be justified in believing that P *via modus ponens*, one must have a good noncircular argument for believing that *modus ponens* is a reliable inference rule.

This courts skepticism about our deductively formed beliefs for reasons that Lewis Carroll pointed out in 1895 in a funny dialogue in *Mind*. Here is my abridged version of the dialogue:

> **Achilles**: I argue that: (i) if A then B, (ii) A, so (iii) B.
> **Tortoise**: [*Clearly impressed with his ingenuity*.] Suppose I accept (i) and (ii), but I don't accept (iii). What would you say to *that*?
> **Achilles**: I say that is unintelligible. You *must* accept (iii) if you accept (i) and (ii).
> **Tortoise**: But why is that so?
> **Achilles**: Because (iii) *follows deductively* from (i) and (ii).
> **Tortoise**: But I don't get why that's a good justification.
> **Achilles**: Well, you goofball, if B follows deductively from A, then if A is true, B *must* be true. In this case, (iii) follows deductively from (i) and (ii), because *modus ponens* is just a case of deductive consequence.
> **Tortoise**: So, what you've just told me is that I should accept the inference from (i) and (ii) to (iii) because *if* (i) and (ii) are true, then (iii) *must be true*. Is that what you are saying?
> **Achilles**: [*Impatiently*] Yes!
> **Tortoise**: But that's *precisely* what I'm rejecting, and you are just presupposing that I'm wrong. I just don't see that if (i) and (ii) are true, then (iii) must also be true. You can't tell me that the reason why I have to accept this inference is that (iii) is a deductive consequence of (i) and (ii) if *all you mean* by "(iii) is a deductive consequence of (i) and (ii)" is that if (i) and (ii) are true, (iii) must be true.

What Carroll's dialogue really suggests is that the JJ Thesis is too strong. If we accept the JJ Thesis, we have to accept JJ-Modus Ponens, since it's just an instance of that thesis. And if we accept this instance, we're put into the position of Achilles in responding to the Tortoise: we have to explain, without simply using *modus ponens*, why *modus ponens* is justified. But it isn't easy to do that! We can't just say (as we might be tempted to say): "*modus ponens* is a deductive rule". For we would have to explain why that is a good reason for using it. And it seems like all we can say about *that* is that <u>if</u> the premises of an argument using *modus ponens* are true, <u>then</u> the conclusion *must be true*. That's all that it *means* to say that *modus ponens* is a deductive rule. But if we were looking for some *independent* justification for *modus ponens*, this is not going to be satisfactory. This simply *presupposes* the correctness of *modus ponens*.

The Tortoise was right: if we were unclear why *modus ponens* was justified to begin with, it's not going to help to note that *modus ponens* is a rule such that if its premises are true, its conclusion must be true, which is all we'd be noting if we insisted that it's a deductive rule.

But this point doesn't show that we can't be justified in using *modus ponens*. It just shows that *modus ponens* is a *basic inference rule*: one for which no independent, non-circular justification in the form of some argument is needed. If there are any basic rules (which there *must* be!), we must reject the JJ Thesis. If we reject the JJ Thesis, we also must reject the inference from (7) to (8).

This takes a lot of the bite out of the Traditional Problem of Induction. It shows that even if we can't *prove with a non-circular argument* that induction is a good method, we might still be justified in using this method. In general, there isn't a problem with this: we also can't (easily) prove with a non-circular argument that *modus ponens* is a good method. The simplest way of trying to do this (i.e., noting that *modus ponens* is a deductive rule, and explaining what that means) presupposes *modus ponens*, and hence isn't a non-circular argument. That was Lewis Carroll's old point.

Still, I think the Problem of Induction should still worry us a little bit. The reason why it should worry us is that there is an obvious disanalogy between deduction and induction. We can easily imagine how induction could go wrong. We cannot imagine how deduction could go wrong. This isn't to say that we've proven that deduction is a great thing: no ordinary person who uses deduction can prove that in the sense of giving a non-circular argument for its reliability. It's just to say that deduction is *obviously* correct, whereas induction is *not* obviously correct. The reasons for thinking that *modus ponens* is a basic rule don't generalize to the case of inductive rules, because it simply *isn't* obvious that these rules are good. (The Tortoise was *silly* for wondering about the quality of a *modus ponens* inference. Hume was *not* silly for worrying about induction.)

It would be nice if something could be said about why induction should be taken seriously as a belief-forming method. But that isn't easy, and *this* is Hume's more modest and more interesting point. Can anything be said about this? Can we answer Hume on his own terms?

Maybe. The only mildly plausible approach I've seen takes the following form. It begins with the observation that induction isn't the only form of non-deductive inference. There is also *abduction*, or *inference to the best explanation*. The difference between the two is clear if we look at their form. An inductive inference has the form:

> All observed Fs are Gs.
> (Implicit Uniformity Principle: observed patterns tend to indicate general patterns that extend into the unobserved parts of the world)
> ---
> So, all unobserved Fs are also Gs.

An abductive inference, by contrast, has the form:

> A striking fact X has been observed.
> The best explanation of X is Y.
> ---
> So, Y is true.

These are very different forms of inference. More importantly, their *prima facie* status as forms of inference is not even close to being equal. It seems *obvious* that abduction is a good form of inference in a way in which it does *not* seem obvious that induction is. How could it be rational to observe some fact, and note that some explanation E is the *best* explanation of that fact, while refusing to accept E as an explanation of that fact? That just seems *incoherent*. Nothing like this

was as clearly true in the case of induction. It *doesn't* seem incoherent on the face of it to be worried about the Uniformity Principle. The Uniformity Principle just isn't as obvious as the principle that we ought to accept the best explanation (if we accept any explanation at all).

This opens up a new line of response to Hume. Hume assumes that we've only got two options for justifying induction: we can appeal to deduction, or we can appeal to induction. The first is no good, because the Uniformity Principle is not a principle of deductive logic. The second is no good, because it begs the question. Hume ignores a third option: we could appeal to abduction to justify the Uniformity Principle.

This is actually a quite reasonable thought. Part of what strikes us as crazy about Hume's counter-inductive hypotheses (e.g., that fire will start freezing us tomorrow, or that food will fail to nourish us tomorrow, or whatever) is that they are much more complicated than the hypotheses we'd ordinarily entertain in attempting to justify our beliefs about the future. It would just be simpler explanation of an observed pattern that all observed Fs are Gs that it's a *law of nature* that all observed and unobserved Fs are Gs. This claim is much more plausible as a lawlike summary of what has happened and will happen than the hypothesis that things are going to radically change tomorrow, so that all unobserved Fs will be non-Gs. It's more plausible because in some sense it is *simpler*, and simpler explanations are better. (It might also be a better explanation for other reasons, but let's set the complications here to one side.)

So, here is one way in which we could try to *abductively* justify the Uniformity Principle. We say:

> Here is a striking fact: all observed Fs have been Gs.
> The best explanation of this striking fact is that a local Uniformity Principle according to which all observed and unobserved Fs are Gs is true.
> ---
> So, such a local Uniformity Principle (i.e., a law of nature) is true.

If we accepted this abductive justification of the Uniformity Principle, we would have induction--independent warrant for the implicit step in every inductive argument. Given justification for that step, the move from the premises of an inductive argument (including the Uniformity Principle) to the conclusion would be rational. This would solve the Problem of Induction.

Of course, this move carries with it some explanatory burdens. While the abductive schema is indeed obviously rational (since it's incoherent to reject the best explanation of some phenomenon, and prefer a different explanation), I've simply assumed without argument that simplicity is something that qualifies an explanation as better. Perhaps this is unproblematic: this too seems obvious. But we can go beyond mere further appeals to obviousness here: the principle that simpler hypotheses are more rational to believe than complicated ones follows from axioms of the probability calculus, and these axioms are themselves basic constraints on rationality. So, while there is an explanatory burden here, I don't think it's a severe one.

In any case, while there is a lot more to be said about the details, I find this to be the most promising approach to solving the Traditional Problem of Induction without simply taking it to be obvious that induction is rational. So, I think we *can* actually confront Hume on his own terms, and not merely rest content with a complacent rejection of the JJ Thesis.

## 2. Popper's Moral

Suppose, however, that you aren't content with this justification, and you also aren't content with a bedrock appeal to the intuitive obviousness of the rationality of induction. Suppose, in short, that you think Hume's problem is insoluble. Would this be a devastating thing to think?

At first, it seems like it would be devastating. After all, don't we think that induction is the necessary foundation of all scientific achievement? Wouldn't Hume's problem thereby undercut the rationality of virtually all scientific belief? It would, if you really thought induction played such an important role in science. One interesting move that some philosophers who are convinced by Hume's argument have made is to simply deny that induction really is the necessary foundation of all scientific achievement. The most famous such philosopher is Karl Popper. Popper thought that when we actually look at what scientists do, we see that they aren't using induction at all. Instead, they are using some combination of abduction and deduction.

Here was Popper's vision. He saw scientists characteristically reasoning as follows:

1. We start with a set of theories $\{T_1,\ldots, T_n\}$ each of which predicts the phenomena we've observed so far.
2. We select the simplest theory from $\{T_i\}$, and *tentatively accept it* for the purposes of research, though without really *fully believing it*.
3. Our task is then to look for experiments which might yield observational data that the simplest theory fails to predict. We try to *falsify* this theory, and also any other less simple theories in $\{T_i\}$.
4. Ideally, we chop down the number of eligible theories in $\{T_i\}$ and continue to tentatively accept the simplest theory that hasn't been falsified, though we stand ready and willing to encounter more falsifying data.
5. Rinse and repeat.

Part of this vision involves never seeing scientists as *believing* anything: what they do instead is *tentatively accept* simple predictively adequate theories, and search for counterevidence. This acceptance is weaker than belief, because it doesn't involve commitment to truth: it just involves a commitment to the theory as a *good starting point* which could probably use improvement. The rest of the vision involves, as I've already said, uses of abduction and deduction. The process of falsification is deductive: we note that a theory *entails* that some observational datum D should occur, then we note that D does not occur, and use *modus tollens* to infer the falsity of the theory. What guides tentative acceptance are abductive considerations like simplicity.

If Popper is right about how scientists actually do their jobs, what he shows is that we can simply *sidestep* the Problem of Induction. If scientists never actually use induction, they shouldn't care if induction can't be justified. Of course, a worry here is whether Popper's empirical claim about scientific practice is true. And it certainly isn't obvious, though he's almost certainly right that *some* scientists live up to his vision. But perhaps what he'd say is that even if they don't live up to his vision, they can, and they ought to: there is simply an alternative to induction that can help us to do science equally well, if not better. This milder moral strikes me as quite plausible.

## 3. The Gettier Problem

One might have gotten the impression that a key difference between science and philosophy is that scientists have "results" and actually converge upon some theoretical truths. And this impression wouldn't be too far off: philosophers love disagreeing with each other, and scientists

do tend to agree in their often quite optimistic tentative acceptance of working hypotheses. But there are some counterexamples to this trend. One of them is the now almost universally held view that the following traditional theory of knowledge is false:

*JTB Theory*: Knowledge is justified true belief.

A striking fact about philosophers' agreement about the falsity of the JTB Theory is that it was brought on by a tiny article published by an otherwise unknown figure named Edmund Gettier.

Gettier's classic paper presents the following two apparent counterexamples to the JTB Theory:

*Who Got the Job*. Smith has very strong evidence for believing the following claim:

(a) Jones will get the job, and Jones has ten coins in his pocket.

The evidence is that Jones showed Smith the ten coins in his pocket a second ago, and that the very sincere and generally reliable president of the company assured Smith that Jones would in the end get the job. Smith infers:

(b) The man who will get the job has ten coins in his pocket.

Smith is clearly justified in believing (b) on the basis of his evidence, and he does believe (b). But, unbeknownst to Smith, he also happens to have, by chance, ten coins in his pocket. Moreover, although the president of the company was being sincere, the company had a last minute change of heart: they decided that Smith, not Jones, will be getting the job. Accordingly, (b) is true.

*The Ford and Barcelona*. Smith has very strong evidence for believing:

(c) Jones currently owns a Ford.

His evidence for (c) is that Jones has at all times in the past owned a Ford, and Smith sees Jones driving a Ford right now. Smith has just taken a logic class in which he learned the following deductively valid rule ("Disjunction Introduction"): *from A, infer A or B*. Smith randomly applies Disjunction Introduction to infer:

(d) Jones currently owns a Ford or Brown is in Barcelona.

Smith has no evidence for the second disjunct in (d): it's just a random claim that he introduced just for the sake of applying the rule of logic he has learned. Still, he is justified in believing (d), because he is justified by strong evidence in believing (c), which entails (d).

As it turns out, unbeknownst to Smith, Jones has sold his old Fords and is planning to buy a Chevrolet. The car he is currently driving is a rental. So, (c) is false. But (d) is true, because, by sheer accident, Brown happens to be in Barcelona.

These apparent counterexamples are widely accepted as decisive: in them, it seems like we have clear cases where someone has a justified true belief that does not amount to knowledge. Still, right after Gettier's paper was published, people were not terribly worried about the credentials of the JTB Theory. They agreed that cases like *Who Got the Job* and *The Ford and Barcelona* were

inconsistent with the letter of the theory, but they thought a simple addition to the theory could save its spirit.

How might the theory be revised? The first thing people noted about Gettier's original cases is that they both involve inference from a false premise. So, people suggested the following view:

> *First Revised JTB Theory*. Knowledge is justified true belief that is not inferred from any false assumptions.

This revision failed to solve the problem. There are cases that have the same moral that falsify the revised theory. Consider:

> *Robot Dog and Real Dog*. Suppose that James, who is relaxing on a bench in a park, observes a seeming dog that is chewing on a bone. So he believes:
>
> (e) There is a dog over there.
>
> Suppose further that what seems to be a dog is actually a robot so perfectly constructed that, by vision alone, it could not be distinguished from an actual dog. James does not know that such robot dogs exist. But in fact a Japanese toy manufacturer has recently developed them, and what James sees is a prototype that is used for testing the public's response. Suppose further that just a few feet away from the robot dog, there is a real dog. Sitting behind a bush, it is concealed from James's view. Accordingly, (e) is in fact true.
>
> *Fake Barn County*. Suppose there is a county in the Midwest with the following peculiar feature. The landscape next to the road leading through that county is peppered with barn-façades: structures that from the road look exactly like barns. Observation from any other viewpoint would immediately reveal these structures to be fakes: devices erected for the purpose of fooling unsuspecting motorists into believing in the presence of barns. Suppose Henry is driving along the road that leads through Barn County. Naturally, he will on numerous occasions form a false belief in the presence of a barn-façade. Since Henry has no reason to suspect that he is the victim of organized deception, these beliefs are justified. Now suppose further that, on one of those occasions when he believes there is a barn over there, he happens to be looking at the one and only real barn in the county. This time, his belief is justified and true.[4]

In these cases, James and Henry don't make any inferences at all. They simply form beliefs *directly* on the basis of visual appearances. Since there is no inference, there is *a fortiori* no inference from a false assumption. Still, intuitively James and Henry fail to know in these cases. Hence the First Revised JTB Theory will be false.

Is there any other way to add a fourth condition to the JTB Theory to get an adequate analysis of knowledge? One useful observation to start with is that what's distinctive about *all* the cases we've considered – and indeed all existing counterexamples to the JTB Theory – is that the beliefs in these cases are *true by luck*. It was sheer luck that Smith also happened to have ten coins in his pocket, and that the company decided at the last moment to change their minds and

---

[4] These two cases were taken almost *verbatim* from the following article by Matthias Steup in the *Stanford Encyclopedia of Philosophy*: http://plato.stanford.edu/entries/knowledge-analysis/#GET.

give him the job. It was sheer luck that Brown was in Barcelona. It was sheer luck that there was a real dog that James didn't notice, and a real barn that Henry happened to stumble upon. All of these points might lead one to think that the following must be true:

> *Anti-Luck JTB Theory*: Knowledge is justified true belief that isn't true merely by sheer luck.

But there is a problem with this suggestion. A theory must be given of what the relevant notion of 'luck' is supposed to be, because there are *other* cases where people *do* know even though their beliefs are true in a lucky way. Consider the following case:

> *Bald Eagle*. You live in a part of the world where bald eagles are extremely rare. By chance and unbeknownst to you, a visitor from abroad brought a bald eagle to your village, and set it free just a few seconds before you decided to look out your window. You look out, see the bald eagle, and say to yourself: "Good heavens! There is a bald eagle over there."

In this case, you form a justified true belief that there is a bald eagle over there. And this justified true belief indeed amounts to knowledge. But it does seem like there is a clear sense in which it is true by luck: it was just sheer chance that someone happened to bring that bald eagle to your country and let it loose just in time for you to see it flying away as you looked outside.

This is not to say that there the Anti-Luck JTB Theory is on the wrong track. After all, it seems important that every counterexample to the JTB Theory does involve luckily true belief. The task is simply to explain what the *relevant* sense of 'luck' needed for the Anti-Luck JTB Theory to succeed might be. We need, in other words, an analysis of the *knowledge-undermining kind of luck*. To date, while there have been many attempts to provide such an analysis, none is accepted as clearly successful. The Gettier Problem remains an open problem.

**MEETING 8**

## 1. Autonomy, Determinism and Some Incompatibilist Pitfalls

### 1.1. *Questions, Distinctions and Views*

In starting our discussion about autonomy and determinism, it is worthwhile to distinguish the following four questions:

1. Does the world really contain *agents*, or individuals who *intentionally cause* events?
2. If the answer to (1) is 'yes', are these agents ever really *responsible* for their acts?
3. If the answer to (2) is 'yes', then what *makes* an agent responsible for her behavior in any relevant case?
4. Does the naturalistic worldview given to us by the hard sciences provide any reason to think that the answer to (1) or (2) is 'no'?

Before we turn to examine a spectrum of views on these questions, notice that they all presuppose a distinction between behaviors that agents *intentionally produce* and behaviors for which agents are *(fully) responsible*, or, almost equivalently, that they *autonomously produce*. That these concepts can come apart might sound slightly counterintuitive at first, but it shouldn't. Consider a case in which a powerful madman threatens to kill your family unless you wound several people. In this case, you might indeed *intentionally* wound these people, but you would not be fully responsible for what you intentionally do, and similarly not be fully free in so acting. Hence, that some act X was intentionally performed by some agent A does not entail that A is fully responsible for intentionally X-ing, or fully autonomous in intentionally X-ing. Indeed, intentional action and autonomous action come apart in precisely this fashion in a large variety of cases: mental illness, inebriation, unwilling addiction, coercion, manipulation, deception, akrasia (= acting against what you believe to be most rational), and so on.

One thing that's worth noting about these cases is that they all intuitively seem to involve a failure to fully reflect the agent's *true self*. If you are temporarily insane, extremely drunk, coerced, manipulated, acting entirely out of an addiction you hate, or acting against your better judgment, you are in some clear sense "not really you", as you would often later recognize. We honor this intuitive point when we say things like: "That's just his addiction talking." This leads to a hypothesis to which we will be returning in due course, which is that a necessary condition for an intentional action to be fully autonomous is that it be one with which the agent *identifies*, that she *endorses*, or that she would regard as *compatible with standards that are crucial to her personal identity*. Let us, however, set aside this hypothesis for a moment and take note of a range of different views that give different answers to the four questions just raised.

Some philosophers hold metaphysical views on which all that *really* exists are the fundamental particles described by microphysics, the intrinsic properties had by these particles, the spacetime manifold in which they are located, and the laws of nature that govern them. Such a view leaves no room for mental properties, and *ipso facto* leaves no room for agents, since the existence of agents clearly requires the existence of some mental properties (e.g., *intending to A*). It is also obviously the case that this view leaves no place for responsible agents. So, this view would give a negative answer to (1) and (2), a positive answer to (4), and would regard (3) as an empty question. While this is an interesting view, we will be largely setting it aside for the purposes of our discussion, since it raises too many peripheral questions.

A different kind of view that is very easily conflated with this view is that the world does contain individuals who intentionally cause certain events, but that the intentional actions of these individuals can nevertheless be explained by scientific laws and lower-level facts, such as facts of neurology, biology and chemistry. This view regards intentional agents as real parts of the world and not just illusions of naïve common sense, but sees all the facts about these agents as *grounded by* lower-level facts described by the "hard" sciences: facts, in turn, which can be explained by laws of nature and the total history of the world.

Notice that, by itself, this different view does not strictly imply a negative answer to (2) or the conclusion that (3) is really an empty question that applies a meaningless concept (i.e., *responsibility*). This view only implies such a further view *given* certain auxiliary assumptions and arguments from them. Nevertheless, many people in the history of philosophy have thought that it is very easy to supply such auxiliary assumptions, and to get a quick argument from the earlier view about what grounds facts about agency to the conclusion that the answer to (2) is 'no' and that there just is no concept of responsibility that has any application.

This leads to a view called *hard determinism*, on which autonomous agency and determinism are *incompatible* (this part is called *incompatibilism*), and on which determinism is true – i.e., on which laws of nature together with the history of the world up until some time *t* determine everything that is true at the subsequent time *t**. The more crucial and debatable aspect of this view is its incompatibilism, and this is what I'll mostly discuss in a critical light.

What exactly are the auxiliary assumptions that, when coupled with a naturalistic worldview that leaves room for the existence of agents, provide an argument for such an incompatibilist picture?

### 1.2. *Incompatibilism: The Argument from the Inability to Do Otherwise*

One of the oldest arguments for incompatibilism takes the following form:

*The Argument from the Inability to Do Otherwise*

A. If determinism is true, then all of my acts are (strict or probabilistic) causal consequences of the events that precede them together with the laws of nature.
B. If all of my acts are (strict or probabilistic) causal consequences of the events that precede them together with the laws of nature, then, in any given case, *I could not have chosen to act otherwise than I actually acted.*
C. If I could not have chosen to act otherwise in some case, I could not have chosen autonomously in that case.
D. So, if determinism is true, I could never act freely; this is just to say that incompatibilism is true.

This argument is now widely believed to be defective, largely due to some important work by Harry Frankfurt that I actually indirectly referenced in earlier meetings. This work is significant enough that I think it is worthwhile quoting the original case that Frankfurt used against (C):

> Suppose someone – [Smith], let us say – wants [Jones] to perform a certain action. [Smith] is prepared to go to considerable length to get his way, but he prefers to avoid showing his hand unnecessarily. So he waits until [Jones] is about to make up his mind what to do, and he does nothing unless it is clear to him ([Smith] is an excellent judge of such things) that [Jones] is going to decide to do something *other* than what he wants him to do. If it does become clear that [Jones] is going to

> decide to do something else, [Smith] takes effective steps to ensure that [Jones] decides to do, and that he does do, what he wants him to do. Whatever [Jones]'s initial preferences and inclinations, then, [Smith] will have his way….
>
> Now suppose that [Smith] never has to show his hand because [Jones], for reasons of his own, decides to perform and does perform the very action [Smith] wants him to perform. In that case, it seems clear, [Jones] will bear precisely the same moral responsibility for what he does as he would have borne if [Smith] had not been ready to take steps to ensure that he do it. It would be quite unreasonable to excuse [Jones] for his action, or to withhold the praise to which it would normally entitle him, on the basis of the fact that he could not have done otherwise….[5]

This case shows that autonomy in A-ing does not entail the ability to do something other than A: in this case, Jones couldn't have done otherwise, since Smith would have interfered and made him A anyway if he hadn't initially wanted to A. A small complaint about cases like this is that, in them, what guarantees that the agent couldn't act otherwise is not something that actually explains his action. You might regard this as a defect of Frankfurt's apparent counterexample to (C), and try to restate (C) as follows:

> (C*) If the fact that someone is unable to do something other than A plays a role in explaining why he does A, then this person couldn't have done A autonomously.

But this would be shortsighted. In later work, Frankfurt notes that in some paradigm cases of autonomous action, an agent's inability to do otherwise plays a crucial role in explaining his behavior. Indeed, somewhat remarkably, it is *precisely* what seems to *make that behavior autonomous* in these cases. Core cases of this kind involve what Frankfurt calls "volitional necessities", which are brought out in the following fascinating quote:

> There are occasions when a person realizes that what he cares about matters to him not merely so much, but in such a way, that it is impossible for him to forbear from a certain course of action. It was presumably on such an occasion, for example, that Luther made his famous declaration: "Here I stand; *I can do no other*." An encounter with necessity of this sort characteristically affects a person less by impelling him into a certain course of action than by somehow making it apparent to him that every apparent alternative to that course is unthinkable….
>
> A person who is subject to [such] volitional necessity finds that he *must* act as he does. For this reason it may seem appropriate to regard situations which involve volitional necessity as providing instances of passivity. But the person in a situation of this kind generally does not construe the fact that he is subject to volitional necessity as entailing that he is passive at all. People are generally quite far from considering that volitional necessity renders them helpless bystanders to their own behavior. Indeed they may even tend to regard it as actually enhancing both their autonomy and their strength of will….
>
> The reason a person does not experience the force of volitional necessity as alien or external to himself, then, is that it coincides with – and is, indeed, partly constituted by – desires which are not merely his own but with which he actively identifies himself…. [Such] volitional necessity may have a liberating effect: when someone is tending to be distracted from caring about what he cares about most, the force of volitional necessity may constrain him to do what he really wants….[6]

---

[5] Frankfurt (1998: 6-7).

[6] Frankfurt (1998: 86-88).

As Frankfurt goes on to point out in a different context, volitional necessities of this sort arguably define the core of a person's identity. This sheds some light on the hypothesis suggested earlier that core cases of non-autonomous intentional action are cases in which a person is "not himself":

> To the extent that a person is constrained by volitional necessities, there are certain things that he cannot help willing or that he cannot bring himself to do. These necessities substantially affect the actual course and character of his life. But they affect not only what he does: they limit the possibilities that are open to his will, that is, they determine what he cannot will and what he cannot help willing. *Now the character of a person's will constitutes what he most centrally is. Accordingly, the volitional necessities that bind a person identify what he cannot help being. They are in this respect analogues of the logical or conceptual necessities that define the essential nature of a triangle. Just as the essence of a triangle consists in what it must be, so the essential nature of a person consists in what he must will. The boundaries of his will define his shape as a person.*[7]
>
> [S]ince [such] necessity is grounded in the person's own nature, the freedom of the person's will is not impaired.[8]

This idea of Frankfurt's provides the most profound objection to the Argument from the Inability to Do Otherwise. What Frankfurt discovered is that in some cases, the inability to do otherwise is *precisely* what makes an act authentically attributable to a person in the sense that makes that act autonomous, and that renders the person responsible for that act.

This, together with Frankfurt's early counterexample, shows the crucial premise (C) in the Argument from the Inability to Do Otherwise – and even the revised version of it (C*) – to be deeply false. There is simply no tight connection between an act's being autonomous and an agent's being able to choose otherwise. Indeed, in cases where an agent is initially bound by her conscience to do some act, and initially finds it unthinkable to do otherwise, it would actually *undermine* her autonomy if she suddenly became weak-willed and found herself easily able to do alternatives to that act which less centrally reflect the core of what she cares about.

In this sense, suddenly acquiring the ability to do otherwise can *deprive* a person of her autonomy. In some more of Frankfurt's lovely and evocative words:

> Unless a person makes choices within restrictions from which he cannot escape by merely choosing to do so, the notion of self-direction, of autonomy, cannot find a grip. Someone free of all such restrictions is so vacant of identifiable and stable volitional tendencies and constraints that he cannot deliberate or make decisions in any conscientious way. If he nonetheless does remain in some way capable of choice, the decisions and choices he makes will be altogether arbitrary…. [A]n excess of choice impairs the will. Without individuality, freedom loses much of its point. The availability of alternatives counts, after all, only for someone who has a will of his own.[9]

As we'll see, this point can be used against some other major arguments for incompatibilism.

---

[7] Frankfurt (1999:114).
[8] Frankfurt (1999: 81).
[9] Frankfurt (1999: 110).

First, though, some further reflections are worth adding to the Frankfurtian theme. Something that I think many of us can concede is that we have no idea why, ultimately, we care so much about the things whose importance to us makes us who we are. I, for instance, haven't the slightest idea how I came to love philosophy so much, or what makes me care so much about the people to whom I devote much of the time in my life. Of course, I do think that philosophy and these people are worth caring about, and so do regard my devotion to them as rationally defensible. But I have no idea what made me think this in the first place; I can certainly imagine that other people wouldn't be inclined to care so much, and can even see that these people wouldn't be irrational in doing so. From my point of view, although my love of philosophy and certain people to whom I am devoted makes sense, I certainly don't see myself as having *chosen* to be the sort of person who would care so much about these matters. Does this make me feel any less autonomous when I then decide to continue writing these notes because I love philosophy, and decide to go home to my fiancée because I love her? It honestly doesn't, even though it's quite clear to me that I chose to love neither of these things.

What this seems to show is that even if we don't choose who we are, and to love the things whose importance to us makes us who we are, all that matters to our acting autonomously is that the mental states that make us act as we do are ones that appropriately reflect who we are. And what this in turn suggests is that *even if* who we are is a matter entirely determined by past events and the laws of nature, as long as *we* in fact exist, our intentional actions may remain autonomous. *This* is the key insight of compatibilism and, of course, precisely what should lead us to see incompatibilism as a red herring. This suggests that, insofar as we think the naturalistic worldview given to us by the hard sciences is compatible with the existence of people like you and me, it is *ipso facto* compatible with these people being autonomous. Any attack on the possibility of autonomous agency must collapse into an attack on the possibility of personhood.

### 1.3. *Incompatibilism: The Consequence Argument*

Let's now see where the second major argument for incompatibilism goes wrong. It is called the "Consequence Argument", and was most famously articulated by Peter van Inwagen:

*The Consequence Argument*

I. If determinism is true, then all events now are (strict or probabilistic) causal consequences of the very distant past together with the laws of nature.
II. No one has, or ever had, a choice about what happened in the very distant past.
III. No one has, or ever had, a choice about what the laws of nature are.
IV. If no one has, or ever had, a choice about A, and B is a (strict or probabilistic) consequence of A, then no one has, or ever had, a choice about B.
V. So, no one has, or ever had, a choice about whether any event that occurs now does occur.
VI. If someone fails to have a choice about whether he does A, then he cannot be autonomous in A-ing.
VII. So, if determinism is true, then no one acts autonomously; this is just to say that incompatibilism is true.

People in more recent, post-Frankfurt literature seem to worry a lot more about the Consequence Argument than about the Argument from the Inability to Do Otherwise. At first, this can seem reasonable. (I) is definitional truth, (II), (III) and (VI) seem obvious, and (V) and (VII) are consequences of earlier claims. Moreover, (IV) *seems* like a truism: surely if you had no choice about A, and A fixedly determines B, then you have no choice about B either.

But I think shouldn't accept (IV), especially after reflecting on Frankfurt's points.

Recall that I conceded that I don't think *I* chose whether I love philosophy so much, or care about my fiancée so much. Moreover, in some cases, I don't think I clearly choose whether my love of these things causes me to do certain things. If someone tried to harm my fiancée, and I was around at the time, I *couldn't but* try stopping them: somewhat like Luther, I would say, "Here I leap to interfere; I can do no other". But my leaping to interfere would be a paradigm case of autonomous action all the same. I would clearly deserve praise if I in fact stopped the assailant, and no one would question that this act was attributable to me.

This suggests that we can easily generate counterexamples to (IV) by applying Frankfurt's theme. We can argue as follows:

i. People do not generally choose to love the things whose importance to them makes them who they are.
ii. Moreover, in some cases, the fact that someone fundamentally cares about something can make it the case that she has no choice *but to act in some way*: she cannot influence whether her deeply caring about something leads her to act.
iii. Some such cases are nevertheless paradigm cases of autonomy.
iv. So, from the fact that someone has no choice about A, and no choice about A's leading to B, it does not follow that someone fails to be autonomous in B-ing.
v. Such cases also seem to be paradigm cases of genuine choice (even though, in them, we couldn't help *but* choose in the way we did). Surely, for instance, I did choose to save my fiancée in the imaginary case considered before. But my love for her still *just kicked in* and made me save her: I can't stop it from doing this (and, indeed, if I could, I would arguably be *less praiseworthy*). (So, while I did choose to save her, I couldn't but choose to do so.)
vi. So, from the fact that someone has no choice about A, and no choice about whether A leads to B, it does not follow that she fails to have a choice in B-ing.
vii. If (iv) and (vi) are true, then (IV) must be false.

When my acting in some way is a consequence of something that I care about very deeply, it is often true that what is causing my act is something which I didn't choose, and indeed that the force with which my caring leads me to act in that way isn't something that I can lessen, and *ipso facto* isn't itself something that I choose. Still, when these kinds of states cause me to act, I act autonomously: my act is attributable to me, and I am responsible for it, and can be praised for it. I even still choose to do this act, though I nevertheless, in a sense perfectly compatible with autonomy, couldn't have helped but choose as I did. Accordingly, the Consequence Argument is no good. There are pretheoretically intuitive cases where (IV) fails.

## 2. What Makes an Act Autonomous?

These points enable us to give optimistic answers to questions (2) and (4) from above: the scientific worldview doesn't seem incompatible with the existence of autonomous action, and so we can preserve our common sense belief that we do sometimes act autonomously. Still, along the way we've strikingly given up certain other (misleadingly) highly intuitive claims about the relationship between autonomy and possibility: we have rejected the claim that autonomy requires the ability to do otherwise, and we have rejected the claim that, if some act can be explained by "unchosen" forces, that act *ipso facto* cannot count as freely chosen. One might begin to wonder, then, how we're going to answer (3) from above. If we reject these (misleadingly) highly intuitive claims about autonomy, what exactly does autonomy amount to?

We can start with the hypothesis flagged earlier. People seem most clearly nonautonomous when their acts fail to fit and appropriately reflect the standards that most deeply ground their identities. Conversely, people seem autonomous in acting to the extent that their acts do fit and appropriately reflect the standards that most deeply ground their identities. This hypothesis seems promising. But it doesn't yet give us a *theory* of autonomy, because it isn't clear what is involved in an act's fitting and appropriately reflecting the standards that ground someone's identity. Accordingly, let's turn to consider some accounts of this.

### 2.1. *Naïve Hierarchical Models of Autonomy*

One of the earliest attempts to pin down this idea was provided – surprise, surprise! – by Harry Frankfurt. Frankfurt noted that one of the things that is distinctive about many of the cases of non-autonomous intentional action with which we began – e.g., coercion, manipulation, deception, addiction, and akrasia – is that, if the agent had reflected on complete information, he would not have *endorsed* the motivational states that led him to act as he did. When a person is coerced to do something, he does indeed end up intending to do it, but he does not endorse this intention: he may wish that he wasn't being coerced precisely because he doesn't really want to intend to do what he is, alas, made to intend to do. Similarly, when someone is willing to do something only because they have been deceived about the nature of what they are doing, they would, once they act and acquire more information, wish that they hadn't had the desires that led them to act. And, even more paradigmatically, when someone is akratic, he by definition fails to act in accordance with his better judgment: on reflection, he would know that what he is doing is not what he ought to do, and would *ipso facto* fail to endorse his act.

This led the early Frankfurt to propose the following model of autonomous action:

> *Naïve Hierarchical Model*: Some intentional action X by an agent A is autonomous if and only if it is produced by motivations within A that A does endorse, or would endorse upon reflecting on full information about X.

Frankfurt in fact proposed a more specific view than the Naïve Hierarchical Model. He followed the dominant trend of assuming the following theory of what it is for someone to act intentionally, a theory which grew out of the work of Donald Davidson:

> *The Desire-Belief Model of Intentional Action* Some agent A does X intentionally iff A desires to X, believes that there is some way W available to him for X-ing, and this desire-belief pair causes him to A in way W.

On this model, the most basic psychological states are desires and beliefs, and intentional agency is just behavior produced in a certain way by these more basic motivating states. If one adopts this general model, how is one going to cash out the notion of *endorsement*? Frankfurt initially cashed it out by saying that someone endorses a desire that he has just when *he desires to have that desire*. Frankfurt's view is called a *second-order desire theory of endorsement*, since a first-order desire is one that has an act as its object, whereas a second-order desire is one that has some other desires as its objects. (The different orders of desire are what lead people to call the resulting models of autonomy "hierarchical models", since these orders create a hierarchy.) So, Frankfurt's more specific early version of the Naïve Hierarchical Model went as follows:

> *Early Naïve Hierarchical Model*: Some intentional action X by an agent A is autonomous only if it is produced by desires of A that A desires to have, or would desire to have upon reflecting on full information about X.

This model leads to some nice predictions about simple versions of coercion, manipulation, deception and akrasia cases. When someone is coerced or manipulated, his act would indeed seem to be produced by motivational states that he would prefer or desire not to have. When someone is deceived in a way that renders his acting non-autonomous, he intuitively would have preferred not to act in that way if he had known more about what he was about to do. When someone is the victim of some addiction (e.g., a smoker), he often desires not to be addicted, but still acts to fulfill his addiction-generated desires. And when someone acts against his better judgment, typically that better judgment *does* cause him to wish that he were acting in a different way: it's just that this wish doesn't exert power over the first-order desires that make him act.

### 2.2. *Problems with Naïve Hierarchical Models and the Importance of Rationality*

Naïve hierarchical models face some objections. One of them is simply that they are unequipped to explain certain intuitive cases of non-autonomy. Recall that, when we started, two of the cases in our original list of non-autonomous acts were acts produced by temporary insanity and acts produced by extreme drunkenness or mental fatigue. It is perfectly possible that an agent might, on appropriate reflection on full information, endorse *some* of the desires he would have when temporarily insane or extremely drunken or mentally fatigued. Consider:

> *The Case of Briefly Crazy Bill*: Bill goes crazy for a bit. In his madness, he desires to eat cereal for breakfast. His reason, however, for desiring cereal for breakfast in his madness is that he thinks that, by eating this cereal, he will cause angels to save the starving children of the world. In fact, Bill ought to eat cereal for breakfast, because it's all he has, and it's good for him. Indeed, when Bill later ceases to be crazy, he looks back on his desire to eat cereal and says: "Wow, I'm so glad I desired to eat that cereal. Otherwise I'd be starving now."

In this case, Bill would end up endorsing his earlier desire to eat cereal, and believe that he was right to desire to eat cereal at that earlier time. Of course, he wouldn't endorse the beliefs that partly led him to have that desire. But he'd still endorse the desire, and that is what matters for autonomy according to the Early Naïve Hierarchical Model. Cases structurally similar to Briefly Crazy Bill can be designed to show the same sort of thing about drunken and fatigued agents.

How might we supplement the theory to handle these cases? Well, one intuitive thing to note about all these cases is that the agent's rational faculties fail to function properly in them. When someone is temporarily crazy, drugged, or fatigued in ways that could relieve him of responsibility for his acts, he will often be reasoning very poorly, or from beliefs that are themselves irrational. This leads to the following revision of the earlier theory, which numerous people in the literature would regard as roughly on the right track:

> *Sophisticated Hierarchical Model*: Some intentional action X by an agent A is autonomous if and only if it is produced by desires and beliefs within A that are formed by properly functioning rational faculties, where these motivational states are ones that the agent does endorse *via* his properly functioning rational faculties, or would endorse *via* such faculties upon reflecting on full information about X.

This Sophisticated Hierarchical Model can handle the autonomy-precluding cases of craziness, drunkenness and fatigue, because, in such cases, the desires and beliefs that lead to the agent's act are often functioning highly improperly. Of course, *a lot* needs to be said about *what* a person's rational faculties are, and what it is for these faculties to function properly. Without a

theory of this, the model either makes no substantive predictions, or makes some bad predictions. I do think a theory can be given, but that's a story for another day.

**References**

Frankfurt, Harry. 1998. The Importance of What We Care About. Oxford: Oxford University Press.

Frankfurt, Harry. 1999. Necessity, Volition and Love. Cambridge: Cambridge University Press.

Van Inwagen, Peter. 1983. An Essay on Free Will. Oxford: Oxford University Press.

## MEETING 9

### 1. Recap, the Failure of Taylor's Argument, and the Irrelevance of Libet's Data

1.1. *Recap*

Last week I presented Frankfurt's fascinating and intuitive reasons for rejecting some crucial assumptions in arguments for incompatibilism, and sketched some elements of his influential version of compatibilism. Given how important Frankfurt's points are (James *really* ought to be discussing them in the lecture!), it is definitely worthwhile to quickly reiterate them. As we'll see, they provide an easy source of resistance to Taylor's argument for his *libertarian* version of incompatibilism, and a straightforward explanation of why Benjamin Libet's neuropsychological discoveries show nothing of importance. (Incompatibilism, remember, comes in two forms. Hard determinists think free will and determinism are incompatible, that determinism is true, and that we thus lack free will. Libertarians agree that free will and determinism are incompatible but say that we have free will and thus that determinism is false.)

The two assumptions that are crucial to the most important arguments for incompatibilism are the following highly general principles about the link between autonomy and possibility:

> **The Principle of Alternative Possibilities**: It is a necessary condition for an agent A to choose and do X autonomously at *t* (or to choose and do X responsibly, or freely, at *t*) that A have the ability to do something other than X at *t*.
>
> **The Transmission Rule**: If it is not up to an agent A whether some state of affairs X obtains, and X's being the case is causally sufficient for some further state of affairs Y, then A has no autonomy over whether Y obtains.

Frankfurt has, I believe, provided a couple of sufficient reasons for rejecting these principles.

One aims mostly at the Principle of Alternative Possibilities, and turns, as you'll recall, on cases like the following:

> *Unused Remote Control.* Jones is completely unaware of Smith and his intentions, but as a matter of fact Smith could make Jones do anything he wanted Jones to do by using a remote control. In fact, if Jones ever wanted to do something that Smith did not want him to do, Smith would use his remote control to make Jones change his mind and do what he wants him to do. As a result, in any given case, Jones cannot act otherwise than he in fact acts, since Smith has a definite opinion on exactly which act Jones ought to perform in any given case. As it turns out, Jones just happens to want to do everything that Smith would want him to do. Accordingly, Smith never has to use his remote control.

This case refutes the Principle of Alternative Possibilities. In this case, Jones could never have acted otherwise than he in fact acted. But, intuitively, all of his acts are still free, since Smith never has to "show his hand", as Frankfurt says.

There is a much more interesting case that Frankfurt has against the Principle of Alternative Possibilities. It is equally a case against the Transmission Rule that played such an important role in the Consequence Argument for incompatibilism to which I exposed you in the last meeting. This case turns on what Frankfurt calls "volitional necessities", which are constraints on what we

are willing and unwilling to do that simply *make us who we are*. These are best understood by pointing again to some of Frankfurt's lovely writings:

> There are occasions when a person realizes that what he cares about matters to him not merely so much, but in such a way, that it is impossible for him to forbear from a certain course of action. It was presumably on such an occasion, for example, that Luther made his famous declaration: "Here I stand; *I can do no other*." An encounter with necessity of this sort characteristically affects a person less by impelling him into a certain course of action than by somehow making it apparent to him that every apparent alternative to that course is unthinkable….
>
> The reason a person does not experience the force of volitional necessity as alien or external to himself, then, is that it coincides with – and is, indeed, partly constituted by – desires which are not merely his own but with which he actively identifies himself…. [Such] volitional necessity may have a liberating effect: when someone is tending to be distracted from caring about what he cares about most, the force of volitional necessity may constrain him to do what he really wants….[10]

> To the extent that a person is constrained by volitional necessities, there are certain things that he cannot help willing or that he cannot bring himself to do. These necessities substantially affect the actual course and character of his life. But they affect not only what he does: they limit the possibilities that are open to his will, that is, they determine what he cannot will and what he cannot help willing. *Now the character of a person's will constitutes what he most centrally is. Accordingly, the volitional necessities that bind a person identify what he cannot help being. They are in this respect analogues of the logical or conceptual necessities that define the essential nature of a triangle. Just as the essence of a triangle consists in what it must be, so the essential nature of a person consists in what he must will. The boundaries of his will define his shape as a person.*[11]

> [S]ince [such] necessity is grounded in the person's own nature, the freedom of the person's will is not impaired.[12]

I think that the idea to which Frankfurt is pointing in passages like this is extremely familiar and intuitive. It's brought out most obviously by cases of the various loves that make us who we are. As I noted in the last class, many of us agree that the following types of claims are true:

(1) We do not generally decide to love the people and things our loving of which makes us who we are. At this very deep level, who we are is not in any interesting sense "up to us", though it is not "beyond us" either, since, well, *it just is who we are*!
(2) Insofar as these loves really make us who we are, we cannot be alienated from them.
(3) In some cases, our loves are so strong, and so important to us, that they make it impossible for us to refrain from acting in some ways, and impossible for us to act in other ways. For example, assuming I'm totally virtuous, I *couldn't but* interfere with some maniac who leaps out clearly trying to harm my fiancée. Moreover, and assuming I'm *minimally* virtuous (as I think I am), I couldn't but refrain from intentionally harming my fiancée. That course of action is *unthinkable* to me.
(4) A consequence of (3) is that I have no higher-level control over whether my love of my fiancée makes me interfere with the maniac, or prevents me from intentionally harming her.

---

[10] Frankfurt (1998: 86-88).
[11] Frankfurt (1999:114).
[12] Frankfurt (1999: 81).

(5) Nevertheless, I am obviously autonomous in interfering with the maniac and refraining from intentionally harming my fiancée. Few acts could more clearly reflect my nature, and could be more centrally what I would identify with.

Claims (1-5) imply the falsity of both the Principle of Alternative Possibilities and the Transmission Rule. The Principle of Alternative Possibilities is deeply false, because when my own nature restricts the course of actions that are available to me to just one (as it does in the case of the attacker, assuming I'm totally virtuous), I am overwhelmingly clearly autonomous. This is one of the clearest cases of self-control imaginable, and is a *paradigm case* of self-direction, which is the concept that actually ***matters***. (Being able to act *arbitrarily* is *not* self-direction: we generally think that if someone really could equally well do just anything (e.g., equally well kill someone as refrain from killing them), he isn't clearly an autonomous person, let alone a person *period*, at all. If freedom were arbitrariness, I believe we would have no reason whatsoever to *care* about whether we're free or not, and this debate would not track any interesting concept.) The Transmission Rule is also deeply false: it's not up to me whether I so deeply love my fiancée – I just *found myself* loving her, and realized that this love is essential to who I am. Moreover, this love necessarily caused me to act as I did in the imagined case. Still, I was *clearly* autonomous in saving her. I obviously would deserve praise in this case.

Cases like this make it seem remarkably clear to me that common sense is in fact compatibilist. Insofar as we really believe there are *selves* in the world – that, among the things that occur in the world, some of them are *us* – we don't in fact see determinism as being incompatible with anything that really matters to us. What matters to us is *self-direction*, not *arbitrariness*. But, if we really *do* believe that some parts of the world are *us*, even if their *being us* is not *caused by us*, we *ipso facto* believe that determinism is completely compatible with self-direction. Compatibilism is in fact the view most clearly suggested by common sense thinking, and the only thing that prevents us from seeing this is, I believe, the tendency to mistake freedom for pure arbitrariness. We don't in fact *care* about pure arbitrariness at all. Most of us would *not* want to be such that it was equally easy for us to kill someone as to refrain from killing them. If "full freedom" actually required this equal easiness, we would cease to care about being fully free. Since we do care about freedom, I think what we *actually* should believe is that it coincides with self-direction, a concept clearly compatible with determinism if we do believe that some parts of the world really are *us*, and really do believe such commonsensical claims as (1-5).

### 1.2. *The Failure of Taylor's Argument and the Irrelevance of Libet's Discoveries*

Taylor fails to see these points. This is excusable: the text in question was written before Frankfurt's views became widely absorbed and understood by the philosophical community.

Taylor's remarkably brisk argument for incompatibilism turns on what he takes to be the two obvious facts that some of our acts are up to us, and that we deliberate. He thinks that we can't deliberate about matters that are not "open to us", and between which we could choose. And he thinks that determinism implies that no possibilities are open to us and so "up to us", and are matters between which we choose. These further claims do not, however, follow from his data.

When who we are constrains the possibilities that we are willing to take seriously in deliberation, and indeed limits the options to just one that we are willing to take seriously, there is, of course, *some* sense in which all other possibilities are not open to us, and *some* sense in which we cannot equally choose between them and the possibility singled out, in effect, by who we are. But that fact shows nothing negative or worrisome about whether the choices that ensue are autonomous, or about whether we in fact deliberated.

Deliberation often consists in a kind of self-discovery: we often have to *learn* what we are unwilling to do and willing to do. Of course, what we are willing and unwilling to do is in fact often a pre-existing matter, because there are strong constraints imposed on our wills by who we are, and who we are is a pre-existing matter. But that doesn't alienate us from ourselves: it just shows us that sometimes we don't know everything about ourselves. If Taylor really reflected on this important truth that deliberation frequently just is self-discovery about the limits of our wills "imposed" by our personal identities, I think he couldn't believe that determinism makes deliberation impossible. *Of course* who we are isn't ultimately up to us. And of course who we are automatically imposes strong limits on what we are willing and unwilling to do. But as long as those limits are part of who we are, we can't be alienated from them, and the fact that, once we reflect, we see that we are only willing to do a certain very restricted number of things doesn't show us that we never make choices. As long as there are definite facts about who we are that are compatible with our being, at bottom, physical things, there really is no problem here.

Of course, you might worry about whether there *are* definite facts about who we are that are compatible with our being, at bottom, physical things. But this is a totally different issue: it's the mind-body problem. Perhaps it is true that common sense is also anti-materialist. But as long as we can believe that mental beings such as *persons* could exist in a purely physical world, I think we can easily be compatibilists. Perhaps what we find here is that the mind-body problem is what we really should be worrying about. This seems to be Thomas Nagel's thought in the following extremely intriguing passage from a paper whose points I'll summarize in a moment:

> I believe that in a sense the problem has no solution, because something in the idea of agency is incompatible with actions being events, and people being things. But as the external determinants of what someone has done are gradually exposed, in their effects on consequences, character, and choice itself, it becomes gradually clear that actions are events and people things. Eventually nothing remains which can be [identified with a] self… [--] nothing but a portion of the larger sequence of events, which can be deplored or celebrated, but not blamed or praised.[13]

I am, however, less skeptical about the possibilities of consistently reconciling the belief we have that persons exist, and that mental beings exist, with the view that everything is ultimately physical. So, as long, I think, as we grant that people could simply *be* certain ways of organizing more basic physical stuffs and things, we should see no problem here.

This, together with the earlier points, may be what shows Benjamin Libet's experiments to establish little of interest to the debate between compatibilists and incompatibilists. Libet's experiments show, at most, that people's intentions are often settled subconsciously before they realize it. (He also reveals the physical basis for this fact. But that, as I've just suggested, implies nothing worrying unless we've already presupposed some anti-materialist view on the mind-body problem.) But if, like I suggest, we view deliberation as a kind of self-discovery, and we grant that undiscovered facts about ourselves can indeed determine what we are willing to do and unwilling to do in a way that isn't at all alienating, this fact shows nothing bad. Don't we *already* believe that we can make genuine self-discoveries, and that we can discover, without becoming alienated and so less autonomous, that unnoticed components of our personality made us unwilling to actuate a large variety of possibilities? I certainly think we do. If so, these scientific results have no import for philosophical debates about (in)compatibilism.

---

[13] Nagel (1979: 37).

## 2. Further Elements of Compatibilism in Ordinary Thinking: Nagel on Moral Luck

So much, then, for an argument that we should have been reading Frankfurt all along. There is another famous discussion that is relevant for establishing a similar conclusion about a further question that remains – namely, Thomas Nagel's discussion of "moral luck". As we'll see, this discussion shows that there is a further way in which most of us already believe essentially compatibilist claims in our ordinary thought about cases.

First, though, let's get this further question on the table. So far I've argued that what we really care about is self-direction and autonomy, and not "freedom" in the quite uninteresting and indeed undesirable sense that means "equal willingness to do absolutely anything". I've argued that compatibilism is true of these more important concepts: determinism does not entail that we lack self-direction or autonomy. As someone astutely pointed out in one of the two sections, one might conceivably insist that *although* self-direction and autonomy are perfectly compatible with determinism, our ordinary moral beliefs may need to be modified in the light of this view. What I now want to ask is to what extent this is really true.

Seeing to what extent this is really true in part amounts to seeing what our ordinary moral beliefs are. In fact, our ordinary moral beliefs are probably inconsistent. (One might reasonably doubt, for example, that anyone can consistently think that it is morally wrong to kill infants and the irreversibly mentally disabled and then eat them while believing that it is morally permissible to kill animals for the negligible difference in pleasure and nutritive value that eating their meat would give us over the often incredibly tasty and nutritious vegetarian alternatives that are available to us (together with dietary supplements). Still, many of us claim to believe both of these things, or act in ways that rationally require us to believe both of these things.) So, *some* of the beliefs to which I am about to point may in fact conflict with other beliefs. All I'm interested in showing for the moment is that some of our beliefs are quite compatibilist when it comes to questions of moral desert, blame and praise.

One thing that our ordinary moral judgments about cases imply is that the following is false:

> *The Anti-Luck Principle*. How much blame or praise some person A deserves in some case cannot turn on matters of luck that are not up to her.

We have beliefs that clearly imply that this principle fails. An obvious case is the fact that we distinguish morally between merely attempted murder and successful murder. We think a successful murderer deserves more blame, and perhaps more punishment, than a would-be murderer whose attempts failed. But note that *whether* a person's attempts succeed can *easily* turn on matters of luck that are not up to her. Some homicidal maniac who likes archery might shoot an arrow on a windy day from the top of a building at someone down below. One gust of wind may blow the arrow off course, but a further, later, gust of wind may return it on track, and may result in the arrow's hitting the person and killing him. If that further gust hadn't come along, this man would have been only an attempted murderer and not a murderer. Yet we blame him more, and punish him more, when he succeeds rather than fails, even though his success in this case is as much a matter of luck as his potential failure would've been. Cases like this show that we often believe the following claim:

> *Pro-Luck Claim*. Due purely to a matter of luck beyond one's control, one can deserve more blame than a counterpart who was not affected by that matter of luck.

If matters beyond our control can make us *more blameworthy*, we clearly do not believe the Anti-Luck Principle, and we also do not believe the characteristically incompatibilist claim that *if* some act's success or failure was not fully up to that agent, that agent would *ipso facto* be at least partly excusable. Criminal law systematically rejects the Anti-Luck Principle when it draws distinctions moral distinctions between things like attempted murder and successful murder. As Nagel notes in his classic paper "Moral Luck":

> Let us first consider luck, good or bad, in the way things turn out…[which] covers a wide range [of cases]. It includes the truck driver who accidentally runs over a child…and other cases in which the possibilities of success and failure are even greater. The driver, if he is entirely without fault, will feel terrible about his role in the event, but will not have to reproach himself. Therefore this example of agent-regret is not yet a case of *moral* bad luck. However, if the driver was guilty of even a minor degree of negligence – failing to have his brakes checked recently, for example – then if that negligence contributes to the death of the child, he will not merely feel terrible. He will blame himself for the death. And what makes this an example of moral luck is that he would have to blame himself only slightly for the negligence itself if no situation arose which required him to brake suddenly and violently to avoid hitting a child. Yet the *negligence* is the same in both cases, and the driver has no control over whether a child will run into his path.
>
> The same is true at higher levels of negligence. If someone has had too much to drink and his car swerves onto the sidewalk, he can count himself morally lucky if there are no pedestrians in its path. If there were, he would be to blame for their deaths, and would probably be prosecuted for manslaughter. But if he hurts no one, although his recklessness is exactly the same, he is guilty of a far less serious legal offense and will certainly reproach himself and be reproached by others much less severely. To take another legal example, the penalty for attempted murder is less than that for successful murder – however similar the intentions and motives of the assailant may be in the two cases. His degree of culpability can depend, it would seem, on whether the victim happened to be wearing a bullet-proof vest, or whether a bird flew into the path of the bullet – matters beyond his control.[14]

As Nagel went on to point out, we believe things like this across the board, and in many very familiar, and completely non-farfetched cases.

*Some* of these cases are, however, structurally very different from the ones just considered. As we've already noted, it is *also* a matter of luck that one turns out to be the sort of person who one essentially is. This can turn on chemical factors and environmental factors that appear at a very early age, and that could easily have turned out very differently. The earlier these factors appear, the stronger their influence can be. When, as compatibilists suggest, these factors become very stable, and it becomes impossible to think of someone as being the sort of person she is without thinking of her as having certain characteristics, we do find it possible to attribute responsibility to her on the basis of the fact that her acts grow out of this essential character. The character was never chosen by her, but she's still responsible, intuitively, for the things that issue from it.

---

[14] Nagel (1979: 29-30).

This is, in effect, the difference between what Nagel usefully calls *constitutive luck* and *resultant luck*. These concepts are even more nicely captured in the Stanford Encyclopedia of Philosophy, which I'll now simply quote:

> **Resultant Luck**. Resultant luck is luck in the way things turn out. Examples include the pair of would-be murderers…as well as the pair of innocent drivers described above. In both cases, each member of the pair has exactly the same intentions, has made the same plans, and so on, but things turn out very differently and so both are subject to resultant luck. If in either case, we can correctly offer different moral assessments for each member of the pair, then we have a case of resultant *moral* luck. [Bernard] Williams offers a case of "decision under uncertainty": a somewhat fictionalized Gauguin, who chooses a life of painting in Tahiti over a life with his family, not knowing whether he will be a great painter. In one scenario, he goes on to become a great painter, and in another, he fails. According to Williams, we will judge Gauguin differently depending on the outcome. Cases of negligence provide another important kind of resultant luck. Imagine that two otherwise conscientious people have forgotten to have their brakes checked recently and experience brake failure, but only one of whom finds a child in the path of his car. If in any of these cases we correctly offer differential moral assessments, then again we have cases of resultant moral luck.
>
> **Constitutive Luck**. Constitutive luck is luck in who one is, or in the traits and dispositions that one has. Since our genes, care-givers, peers, and other environmental influences all contribute to making us who we are (and since we have no control over these) it seems that who we are is…largely a matter of luck. Since how we act is…a function of who we are, the existence of constitutive luck entails that what actions we perform depends on luck, too…. [I]f we correctly blame someone for being cowardly or self-righteous or selfish, when his being so depends on factors beyond his control, then we have a case of constitutive moral luck. Further, if a person acts on one of these very character traits over which he lacks control by, say, running away instead of helping to save his child, and we correctly blame him for so acting, then we also have a case of constitutive moral luck.[15]

How should a compatibilist view our ordinary beliefs about moral responsibility in these two different types of cases?

Well, one thing that people like Frankfurt *must* say is that constitutive luck is a type of moral luck with which we're going to have to live. It is, however, compatible with this variety of compatibilism to adopt a more revisionary proposal about resultant luck. In that kind of case, the luck owes not to one's own nature, but to circumstantial factors that are in no way related to one's essential nature. In cases like this, it is tempting to think that people ought only to be blamed and punished for the factors that did not causally issue from them in any way. Consistently applying this idea, which ends up conceding a little bit to the incompatibilist about moral responsibility without conceding his entire theory, would lead to a revision of many of our ordinary moral beliefs about blame, praise and punishment.

The hardest question is going to be *how* we figure out *what* any given person really essentially is at any time, and how we can isolate his true self and what events are outgrowths of it at

---

[15] http://plato.stanford.edu/entries/moral-luck/#3

some time from various entirely circumstantial factors. As Nagel suggests in his characteristically gorgeous prose style, this problem is really hard:

> The [hardest version of the] problem arises, I believe, because the self which acts and is the object of moral judgment is threatened with dissolution by the absorption of its acts and impulses into the class of events. Moral judgment of a person is judgment not of what happens to him, but of him. It does not say merely that a certain event or state of affairs is fortunate or unfortunate or even terrible. It is not an evaluation of a state of the world, or of an individual as part of the world. We are not thinking just that it would be better if he were different, or did not exist, or had not done some of the things he has done. We are judging *him*, rather than his existence or characteristics. The effect of concentrating on the influence of what is not under his control is to make this responsible self seem to disappear, swallowed up by the order of mere events.[16]

If we had a principled account of how to separate people from their environment, and to identify their essential nature without collapsing them, as Nagel says, into the order of mere events, we would *ipso facto* have a distinction between resultant luck and constitutive luck. Compatibilists should look for such a principled account, preserve our commonsense commitment to the possibility of constitutive moral luck, and probably go revisionary to *some* degree about our commonsense commitment to the possibility of resultant moral luck. So, compatibilism of the most plausible kind – Frankfurt's kind – does end up conceding a *bit* of the revisionism suggested by incompatibilist reflections on the chanciness of our acts.

The remaining question for sensible compatibilists will be to sort out their account of what most essentially makes us who we are, since the best compatibilists tie autonomy to self-direction, and view self-direction as action that is a product of the mental states that reflect who we really are.

Whether this question can be answered is unclear to me. But this isn't exactly a strike against compatibilism. Compatibilists may be *right* that the kind of freedom we care about and desire to have is self-direction in this special sense, and that, in principle, this kind of self-direction is compatible with determinism. They may simply be too optimistic about whether this kind of self-direction could ever occur for the quite *different* reason that reflecting on the fact that we are also, at bottom, physical objects reveals that we are less clearly distinctive and identifiable entities than we ordinarily believe ourselves to be. But we shouldn't spurn compatibilism because this different problem is incredibly difficult. (It may be the hardest problem in philosophy!)

**References**

Frankfurt, Harry. 1998. The Importance of What We Care About. Cambridge: Cambridge University Press.

Frankfurt, Harry. 1999. Necessity, Volition and Love. Cambridge: Cambridge University Press.

Nagel, Thomas. 1979. Mortal Questions. Oxford: Oxford University Press.

Williams, Bernard. 1981. Moral Luck. Cambridge: Cambridge University Press.

[16] Nagel (1979: 36).

## 1. Personal Identity and What Does and Does Not Matter

### 1.1. *Three Questions and the Difference between Qualitative and Numerical Identity*

It is crucial to distinguish at least three different questions in discussing what often comes under the unfortunately broad heading of "personal identity" in contemporary philosophy:

1. What determines whether some individuals X and Y at two different times *t* and *t** are numerically the same person?

2. What aspects of a person at some single time are essential to the distinctive self with which she would identify most deeply?

3. What ought to matter to us in contemplating our persistence over time?

The answers to these questions are importantly separable. Their significant differences become clear when we reflect on imaginary but certainly possible cases like the following:

> *Three Options*. A powerful madman kidnaps you and offers you a three-way choice.
>
> On the one hand, he could give you $3,000 if you agree to be put into the Experience Machine for a day. If you agree to be put into the Experience Machine, most of your deepest beliefs, desires, loves, intentions and goals will end up changing vastly as a result of your interaction with the peculiar virtual reality that it presents. Indeed, you will emerge being quite disdainful of your earlier self and very grateful for having entered the Experience Machine – not just because you will have more money, but because you will come to think – perhaps wrongly! – that you have *better* beliefs, desires, loves, intentions and goals. But you can't know in advance how you will change: all you know is that the changes will be massive.
>
> On the other hand, he could steal all the money from a bank account in which you happen to have about $1,500, but leave you exactly as you are without harming you in any other way.
>
> Finally, he could painlessly kill you but create a clone of you that would go on to interact with your friends, professional relations, lovers, et al., exactly as you would have done, and to whom he will give $3,000.

One of the interesting facts about *Three Options* is that most of us do not regard the first and third options as being remotely similar in choiceworthiness. Most of us would strongly prefer the first option over the third – though, of course, most of us would *ultimately* pick the second. This shows that, on one level, we *appear* to care fundamentally about *numerical* rather than *qualitative* identity. After all, the clone will be *qualitatively just like you*, and will be, by almost everyone's lights, a fully convincing successor. But the future self that will emerge from the Experience Machine will be *qualitatively vastly different* from your current incarnation.

*Why* is it rational to prefer the first option to the third? One might offer the following reasonable answer: "While I would change a lot if I picked the first option, at least I wouldn't *die*. I would clearly *cease to exist* if I picked the third option. Surely I'd prefer to continue to exist, even if I were to be deeply changed, than to die." And this would be superficially plausible – though, as we shall later see, partly deeply mistaken. To be sure, we can say things like: "Since his marriage, Jones has been a different man". But in saying this, we don't suggest that marriage is death. So, we in general think that question (1) is quite different from question (2): what

makes a person continue to exist as a numerically selfsame individual over time is different from what a person's distinctive, deeper self at some given time might be. What *seems* to justify us in preferring the first option to the third is that we would *live on* (albeit highly changed) in the first, whereas we would be *dead* in the third (albeit replaced by a qualitative duplicate).

In this way, numerical identity might appear to matter more to us than qualitative identity. But while, on one level, we do appear to care about numerical identity, since it seems to be what determines whether we *survive* rather than *cease to exist*, we still care about qualitative identity. This is made clear by the fact that many of us would *not* prefer the first option over the second in *Three Options*. We would prefer not to be drastically qualitatively altered – e.g., be forced to come to have vastly different core beliefs, goals, intentions, plans and loves – even if we would get a big material benefit. Indeed, we would prefer this greatly enough to suffer a material loss.

Accordingly, the answers to questions (1), (2) and (3) can come apart in ways that one might not have pre-reflectively anticipated.

In the case of (1) versus (2), we believe we could continue to exist – i.e., *survive* rather than *die* – even if we underwent a vast qualitative change, and indeed lost many of the desires, intentions, beliefs, plans and loves with which we most deeply identify. As it happens, this type of vast change has surely befallen many of us. Most of us are qualitatively very different from ourselves at age 10. We have strikingly different beliefs, aims, and conceptions of what we ought to value and of "what we want to be when we grow up". For some of us, if we could imagine ourselves going back and telling our ten year-old selves what we are like now, those earlier selves might be disgusted, contemptuous, outraged, and unable to accept the prospect of becoming who we now are. We might, at that age, have wanted deeply to become athletes or computer programmers. If we had contemplated the thought that we might come to find these professions uninteresting or silly, and might come to deeply love careers as scholars of Elizabeth Bishop or as philosophers, we might have then laughed. But not now! Still, if that's what happened, we do not *literally* say: "What *really* happened is that the earlier person *died*." We say: "That was me, all right: thank heavens I've changed so fundamentally!" In this way, two qualitatively distinctive selves might exist at different times and yet be clearly parts of one and the same person.

When it comes to the third question, what we find is a nuanced and complex answer that demands prior answers to both (1) and (2). In one respect, it seems like what we want is to *survive*, and to exist as *numerically the same individual* over a long span of time. In another respect, we deeply value the qualitative characteristics that happen to make up who we most deeply are at a time *while acknowledging* that we may survive as numerically the same person even while coming to acquire a very different distinctive set of values, beliefs, goals, intentions, loves, and so on.

### 1.2. *Is Numerical Identity Really Important?*

What ought to matter most to us? There is a fascinating argument worth mentioning to the effect that we ought to forget about numerical identity, and care only about a particular kind of materially sustained qualitative identity. Consider:

> *Division.* Your brain happens to be constituted in such a way that all the important information in it is redundantly encoded in both hemispheres. As a result, if you lost one hemisphere, this would have no effect on your mental life. A crazy but extremely skilled brain surgeon drugs you into a deep sleep, removes your brain while sustaining its functioning by external support, and splits the hemispheres. Now, you also happen to have had two qualitative twins. The surgeon removes and

> destroys their brains and uses their bodies as the new "houses" for your two hemispheres. He finishes, and two individuals with your mental life now awaken.

If you knew that this was going to happen, it is very hard to see why you should be upset in the *same way* in which you should if you learned that someone was going to simply kill you by shooting you in the head. After all, consider your intuitions about the following case:

> *Partial Destruction*. Your brain happens to be constituted in such a way that all the important information in it is redundantly encoded in both hemispheres. As a result, if you lost one hemisphere, this would have no effect on your mental life. A crazy but extremely skilled brain surgeon drugs you into a deep sleep, removes one of the hemispheres of your brain, and destroys it. He then removes the other hemisphere and places it into a different, perfectly healthy body. As it happens, it happens to be the body of one of your qualitative twins; the surgeon has removed this twin's brain, and has used the old body as the new "house" for your remaining hemisphere. He finishes, and you wake up, not able to detect any difference.

There is only one difference between *Partial Destruction* and *Division*: in *Division* but not *Partial Destruction*, both hemispheres continue to exist. If you have no reason to regard what happens to you in *Partial Destruction* as being as bad as what would happen to you if you were simply shot in the head and killed, it seems like you could hardly have a reason to feel differently about *Division*.

But if this is right, it follows that surviving as numerically one and the same person cannot *fundamentally* matter to us. After all, there is an extremely simple argument that you really *do* numerically cease to exist in *Division*. Consider this seemingly airtight reasoning:

i. One individual cannot be numerically identical to two numerically distinct individuals. [This is a logical truth.]
ii. The two individuals who wake up at the end of *Division* are numerically distinct. [This is an obvious fact.]
iii. So, you cannot be numerically identical to both of them. [This is a consequence of (i) and (ii).]
iv. But you also cannot be numerically identical to either one of them. There would be no reason for claiming that one is numerically identical to you that isn't also a reason for claiming that the other is numerically identical to you: each bears the same physical and psychological relations to your original self as the other. It would be *arbitrary* to say that you are one but not the other. [This is also an obvious fact.]
v. If claims (i) through (iv) are true, then, at the end of *Division*, there is no individual that is numerically identical to your earlier self. [This is also obvious.]
vi. If, at $t+$, there is no individual that is numerically identical to some earlier individual X that existed at $t$, then X no longer exists at $t+$. [This is another logical truth.]
vii. So, since claims (i) through (iv) are true, you no longer exist at the end of *Division*. [This is an obvious consequence of earlier claims.]

Logic forces us to claim that you cease to exist in *Division*: you cannot survive as both, but it would be arbitrary to claim that you survive as only one, so the only conclusion is that, *numerically speaking*, you survive as neither. Still, it seems overwhelmingly intuitive to think that you should not be as afraid of the prospect of the type of surgery that occurs in *Division* as you should be of the prospect of someone simply killing you in an ordinary case by, say, shooting you in the head.

It follows from all this that surviving *in the sense* of having some future counterpart that is numerically identical to your current self isn't, most fundamentally, what matters.

What matters would seem instead to be something like this: having a successor who is psychologically continuous with you, and whose mental life is sustained by some of the same physical material that sustained your earlier mental life. Of course, in most ordinary cases, this simply suffices for numerical identity. But, as we see in cases like *Division*, it is not generally sufficient for numerical identity. Since we also see that the loss of numerical identity in cases like *Division* cannot be rationally regarded as being as worrisome as death in the ordinary sense, we also see that the preservation of numerical identity isn't really of any importance *by itself*.

### 1.3. *Some Dimensions of What Really Matters*

These types of considerations suggest a mixed view that incorporates both a psychological and a physical criterion of personal identity. We can hold what seems to be the very attractive

> *Hybrid Theory*, on which X at *t* is the same person as Y at *t** iff (i) Y is psychologically continuous with X, (ii) the physical basis for Y's mental life is at least partly the same as the physical basis of X's mental life, and (iii) there is no other mental life Z sustained by at least partly the same physical basis of X's mental life that is *as* psychologically continuous with X's mental as Y's mental life is.

In cases where there are not multiple candidates for successors of X, as there are in *Division*, the relation described by the Hybrid Theory will be the relation of numerical personal identity. When there are multiple candidate successors, numerical identity ceases to matter; what we should care about is having *some* successor that's psychologically continuous with us, and whose mental life is sustained by at least partly the same material basis as our earlier mental life.

The Hybrid Theory appeals to the notion of psychological continuity. How exactly should this notion be analyzed? We cannot, after all, analyze it in any arbitrary way: some ways of understanding psychological continuity would make this theory turn out false.

Well, we've seen some suggestions about how to understand this relation in the lecture. On a Lockean view, the most important type of psychological link for understanding personal identity is the link created by *apparent memories*. According to what we can call

> *The Simple Memory View*, two individuals A and B existing at distinct times *t* and *t** are perfectly psychologically connected iff they share all the same apparent memories, and psychologically connected *to some degree* N *to the extent* that they share apparent memories (e.g., N = 0 if they share none, N = .5 if they share half, N = 1 if they share all, and so on).

Of course, mere connectedness cannot be the whole story. We can imagine a series of cases over a long span of time: in case 1 at time *t*, you lose a memory of a time before *t*, in case 2 at *t**, you lose another memory of a time before *t*, in case 3 at *t*** you lose another memory of a time before *t*, you lose yet another memory, and so on. All the while, you continue to accrue new memories of things that happened after *t*. If the series is extended long enough, eventually your later self will share none of the apparent memories had by your earlier self at *t*. Still, you could intuitively easily remain numerically the same over time. The *continuity* relation that we need isn't *connectedness*, but something that can be defined in terms of connectedness. We can uphold this

> *Definition of Psychological Continuity*: Two individuals A and B existing at distinct times *t* and *t** are psychologically continuous iff there is some series of person-stages $X_1$,…, $X_n$ such that A is strongly psychologically connected to $X_1$, $X_1$ is strongly psychologically connected to $X_2$, $X_2$ is strongly psychologically connected to $X_3$, … and $X_n$ is strongly psychologically connected to B.

So understood, the series case I just discussed is one in which psychological continuity is maintained, assuming, as the Simple Memory View suggests, that apparent memory links are the most important links. In that example, the members in each pair of person-stages in the series were strongly psychologically connected. It's just that, as we went step by step, little by little was lost from what was present in the first stage: this was, however, perfectly consistent with a high degree of *pairwise* connectedness. Given that there was strong pairwise connectedness in every case, our definition of continuity allows that the first and last members of the series were psychologically continuous. And this, as we wanted to say, was what really played a determinative role in numerical personal identity.

This is a step forward. But the Lockean idea of cashing out psychological connectedness and hence psychological continuity *solely* in terms of apparent memory connections cannot be the whole story. There are *many* dimensions of psychological similarity. Of two person-stages, we can ask: (i) would these stages share similar *beliefs*, (ii) would these stages share similar *desires*, (iii) would these stages share similar *intentions*, *goals* and *plans*, (iv) would these stages share similar *loves*, and (v) would these stages share similar *character traits* (e.g., amicability, open-mindedness, steadfastness, etc.)? All of these dimensions partly influence the extent to which we are willing to claim that two adjacent stages really are stages of some single person. If, over the course of one second, some mad neuroscientist changed all of a person's beliefs, desires, intentions, goals, plans, loves, and character traits, and had these irreversibly replaced by completely opposing beliefs, desires, intentions, goals, plans, loves, and character traits, it would be very hard to claim that the resulting person really is numerically the same person. If it was ever intuitive to claim that psychological continuity played *some* role in determining whether some single person persists over time, we surely wouldn't want to claim that similarities in apparent memory are *all* that matter. This simply wouldn't be a complete story.

So, we need to replace the Simple Memory View by something like

> *The More Complete View*, which holds that two individuals A and B existing at distinct times *t* and *t** are perfectly psychologically connected iff they share all the same apparent memories, beliefs, intentions, goals, desires, plans, loves, character traits, and so on, and are psychologically connected *to some degree N* to the extent that they share such psychological features (e.g., N = 0 if they share none, N = .5 if they share half, N = 1 if they share all, and so on).

Given the More Complete View of psychological connectedness and the notion of psychological continuity that results from substituting the content of this view into our earlier Definition of Psychological Continuity, we get the basis for a more plausible interpretation of the Hybrid Theory. Is the theory that results good enough? I think it isn't, for reasons that should be clear from our earlier discussion of Frankfurt-style views of autonomy.

Recall that, on the accounts of autonomy that grow out of Frankfurt's work, a person acts autonomously when she acts in ways that appropriately reflect her true self, and the standards that make up that self. On such accounts, not just *any* desire or intention or character trait that

you have is really part of your true self or reflects the standards that make up that self. If someone is addicted to some substance, but hates this addiction and wishes that he didn't have it, we want to separate the person from the desires that are generated by the addiction, and claim that the behavior that ensues from these addicted desires which the person doesn't endorse in some sense "isn't really his behavior". The intuitive force of this thought is, as I noted, reflected in familiar claims from ordinary life like, "That's just your addiction talking, not you." These claims could, as we suggested in line with Frankfurt, be more than merely metaphorically apt: they suggest that we tend to identify a person's deeper self with the desires with which that person actually *identifies*, or would identify after an appropriate amount of rational reflection.

Now, you might reject Frankfurt-style accounts of autonomy. But there was surely *something* right in these theories: there is clearly an asymmetry between psychological attitudes that a person *reflectively rejects* and psychological attitudes that she *reflectively endorses*, or would reflectively endorse after an appropriate amount of reflection. I think we want our account of personal identity to reflect this asymmetry. As it stands, the version of the Hybrid Theory that results from endorsing the More Complete View of psychological connectedness does not do so.

After all, the More Complete View says that it is a necessary condition for full psychological connectedness that two person-stages share *all* the same apparent memories, beliefs, intentions, goals, plans, desires, loves, and so on. This clearly fails to honor the significant distinction between reflectively rejected intentions, goals, plans, beliefs, desires and so on, and reflectively endorsed states of these types. What we want, then, is

> *The Yet More Complete View*, on which an individual B existing at some time $t^* > t$ is psychologically connected with the earlier individual A at $t$ in the *relevant sense* to the extent that B retains A's apparent memories, and also retains the beliefs, intentions, goals, desires, plans, loves, character traits, and so on, that A would not have reflectively rejected, or that A would have reflectively endorsed.

When we plug in the notion of *relevant* psychological continuity that results from the Yet More Complete View and the Definition of Psychological Continuity in terms of *relevant* psychological connectedness, we get a more plausible version of the Hybrid Theory. After all, we would not have wanted to claim that if some guy is addicted to some substance at $t$, and *hates* this addiction at $t$ and wishes desperately that he did not have it but then, by some miracle, loses the addiction at a later time $t^*$, he is *less his earlier self*.

Of course, if we made the mistake of identifying the man with his addiction, we might claim this. But when we better understand what it takes for some set of desires and inclinations to be reflective of someone's true self, I think we would no longer want to make this claim, at least if the man was truly wholehearted in his hatred of his addiction and in his strong (but ineffective) desire that it dissipate. Yet notice that if the man in question was truly strongly addicted at $t$, his desires, intentions, goals, and plans at $t^*$ may be *substantially different*. If we hadn't distinguished between mere psychological similarity and the relevant concept of psychological connectedness, and honored this distinction in our theory, we would have to claim that *simply* because these desires, intentions, goals, and plans are very different at $t^*$, the person-stage that exists at $t^*$ is not as good of a candidate for being this man's numerically selfsame successor as it would have been if he had kept the addiction. We would have had to claim that a person-stage that retained the addiction and all the desires, plans, inclinations and intentions that were bound up with it would have been a *better* candidate for that man's numerically selfsame successor. But this would have been a mistake: we want to say exactly the reverse.

In this way, many simple theories of personal identity that appeal to *mere* psychological connectedness and a related notion of *mere* psychological continuity are not sufficiently nuanced. Sometimes a superficial dissimilarity between two person-stages should *not* compel us to find it *less plausible* that these are parts of the numerically selfsame person. If the superficial dissimilarity between the later stage and the earlier stage is generated by a loss of desires, intentions, inclinations, intentions, goals and plans which the earlier stage decidedly did not endorse, and indeed that he strongly wished at a reflective level not to have, this dissimilarity is not one that should be relevant to our view about whether these stages bind together in a single someone.

To put it in a brief slogan: if you change in ways in which you deeply want to change, you are a good example of *self-actualization*, not of *partial self-destruction*!

## 2. Weighing the Dimensions: Material vs. Psychological Continuity

A lot more could be said about exactly how the account of relevant psychological continuity should be tweaked to leave us with a maximally plausible incarnation of the Hybrid Theory. I'll save this for our next meeting. For now, it's worth discussing at more length exactly how we ought to weigh the two dimensions that figure in the Hybrid Theory. Here I'll want to open things up for discussion, since I am myself somewhat uncertain what to say about this issue.

It is possible to get intuitions to pull in two different directions. On the one hand, when we adopt a third-person point of view, we certainly want to say that if at the end of someone's life, that person ends up with an extreme case of Alzheimer's and begins acting in a radically childlike way, the person hasn't literally ceased to exist: she isn't yet *dead*. Yet the psychological discontinuity between this stage of the person and the earlier stages may be great. Moreover, the onset of this discontinuity may be very sudden. (This might be more obvious in a case in which someone sustained a severe blow to the head that resulted in irreversible amnesia: just imagine the case in the way that makes the intuition most compelling for you.) If we continue to believe that the person continues to exist, this can make it seem as if the part of the Hybrid Theory that emphasizes the continuity of the matter that sustains the person's mental life[17] is the most crucial aspect of the persistence of some numerically selfsame individual over time.

At the same time, when we adopt a first-person point of view, intuitions pull in a different direction. Plenty of people write *living wills* or *advance directives* that instruct caretakers fail to do certain things under the condition that their later "selves" completely lose psychological continuity with their earlier selves. Someone might write in a living will: "If I contract a terminal illness when I become severely and irreversibly demented, please do not provide me with life support." Why do people's first-personal attitudes change so significantly when they contemplate the prospect of the kind of radical psychological discontinuity brought on by severe dementia or Alzheimer's? People literally seem to regard this prospect as analogous to *death*. If they are justified in regarding it in this way, it would seem that the mere existence of the same material basis that sustained the earlier mental life is not really relevant at all.

[17] This would, however, have to be *together* with the constraint that this matter is still sustaining *some* mental life, even if it's very different from the earlier mental life. On reflection, few of us would want to claim that somebody who entered a truly irreversible coma or permanent vegetative state was still an existing person: the *person* ceased to exist, though the *body* lived on.

## Why We Cannot Accept Purely Psychological or Physical Views of Personal Identity

### 1. *Where We Stood at the End of the Last Meeting and Some Comparisons with Parfit's View*

In the last meeting, I got a bit ahead of myself and quickly suggested that we adopt a

> *Hybrid Theory*, on which X at *t* is the same person as Y at *t** iff (i) Y is psychologically continuous with X, (ii) the physical basis for Y's mental life is at least partly the same as the physical basis of X's mental life, and (iii) there is no other mental life Z sustained by at least partly the same physical basis of X's mental life that is as psychologically continuous with X's mental as Y's mental life is.

I then went into a lot of detail about how exactly we ought to understand the relation of psychological continuity. I left us, however, with a puzzle. As I suggested, it is easy to get intuitions to pull in two opposing directions, suggesting either that the psychological part of the Hybrid Theory is irrelevant or that the physical part of the Hybrid Theory is irrelevant. Today I want to provide a somewhat indirect way of seeing these opposing intuitions as failing to be objections to the Hybrid Theory. This will just involve showing why we cannot accept either a purely physical theory or a purely psychological theory. In doing this, I'll end up reviewing some of the key arguments from Thomson and Parfit. Thomson holds an implausible purely physical theory for insufficient reasons but has interesting objections to purely psychological theories and also to hybrid theories. Parfit, on the other hand, offers some nice ways of showing that her objections to these theories fail.

Before going into this, it's worth noting that the Hybrid Theory I stated is *close* to the view that Parfit recommends. But what Parfit calls "[his] view" is not really a complete theory of personal identity. It instead consists simply in the following two claims, which I take *verbatim* from his paper:

> 1. If there will be a single future person who will have enough of my brain to be psychologically continuous with me, that person will be me.
>
> 2. If some future person will be neither psychologically continuous with me, nor have enough of my brain, that person will not be me.[18]

It is more useful to state the second claim in the following logically equivalent form:

> 2*. Some future person will be me *only if* that person is *either* psychologically continuous with me *or* will have enough of my brain.

(2*) expresses a *necessary* condition for numerical personal identity over time: it says that it's *required* for some future entity to be numerically identical to some earlier existing person that it either have "enough" of the same brain as the earlier person or be psychologically continuous with that person. (1), on the other hand, expresses a *sufficient* condition for numerical personal identity over time: it says that it's sufficient for some future entity to be numerically identical to some earlier existing person that it have enough of the same brain to sustain psychological continuity with the earlier person. These conditions do not yield *full* necessary and sufficient conditions for numerical personal identity, since there are some cases that they jointly fail to classify. Some examples are cases where a later entity is psychologically perfectly continuous with an earlier person, but where this entity's psychology is transplanted into a new brain. This

[18] Parfit (2008: 178).

satisfies the necessary condition expressed by (2) and the equivalent (2*), since this later entity is indeed *either* psychologically continuous with the earlier person *or* in possession of enough of its brain, for the simple reason that, well, it's psychologically continuous with the earlier person. Moreover, since (1) is *only* a sufficient condition, it has nothing to say about this case at all. So, Parfit's view strictly speaking fails to tell us what happens in this kind of case.

Parfit seems to be happy to leave some cases open, and to have an incomplete view in the sense that it fails to classify all cases. To some degree this is reasonable, since it allows us to say less radical things about certain cases. Consider, for instance, how last week I argued that in the following type of case, we must claim that the individual ceases to exist, and that this reveals simply that numerical personal identity isn't really what matters:

> *Division.* Your brain happens to be constituted in such a way that all the important information in it is redundantly encoded in both hemispheres. As a result, if you lost one hemisphere, this would have no effect on your mental life. A crazy but extremely skilled brain surgeon drugs you into a deep sleep, removes your brain while sustaining its functioning by external support, and splits the hemispheres. Now, you happen to have two twins. The surgeon removes and destroys their brains and uses their bodies as the new "houses" for your two hemispheres. He finishes, and two individuals with your mental life now awaken.

The argument that you cease to exist in this case was simple: (i) you can't be numerically identical to both of the two future persons, since they are numerically distinct, but (ii) you also can't be numerically identical to only one of them, since such a claim would be *arbitrary*: any reason for claiming that you are one is *ipso facto* a reason for claiming that you are the other; so, (iii) you are numerically identical to neither, and so (iv) you cease to exist, since you would continue to exist only if there were some future entity that is numerically identical to you, and there isn't.

Parfit himself was the inventor of this type of case.[19] What did *he* say about it? What he originally said is subtle and a little misleading:

> There will be two people, each of whom will have the body of one of my brothers, and will be fully psychologically continuous with me, because he has half of my brain. *Knowing this, we know everything.* I may ask, 'But shall I be one of these two people, or neither?' But I should regard this as an empty question. Here is a similar question. In 1881 the French Socialist Party split. What happened? Did the French Socialist Party cease to exist, or did it continue to exist as one or the other of the two new Parties? Given certain further details, this would be an empty question. Even if we have no answer to this question, we could know just what happened.
>
> I must now distinguish two ways in which a question may be empty. About some questions we should claim both that they are empty, and that they have no answers. We could decide to *give* these questions answers. But it might be true that any possible answer would be arbitrary. If this is so, it would be pointless and might be misleading to give such an answer….
>
> There is another kind of case in which a question may be empty. In such a case this question has, in a sense, an answer. The question is empty because it does not describe different possibilities, any of which might be true, and one of which must be true. We could know the full truth about this outcome without choosing any of these descriptions. *But, if we do decide to give an answer to this empty question, one of*

[19] See Parfit (1984: 254 – 266) for the original discussion of *Division*.

> *these descriptions is better than the others. Since this is so, we can claim that this description is the answer to the question. And I claim that there is a best description of the case where I divide. The best description is that neither of the resulting people will be me.*[20]

This passage may seem to be in tension with itself. At the beginning, Parfit claims that in knowing *simply* that "[t]here will be two people, each of whom will have the body of one of my brothers, and will be fully psychologically continuous with me, because he has half of my brain", we "*know everything*". But in saying that this is *everything* there is to know, Parfit may seem to be saying that there is *no further fact of the matter* to be discovered. But he goes on to claim that there *is* a further fact of the matter: namely, that the best description of this case is as one in which neither of the resulting people would be me, and *ipso facto*, by simple logic, in which I cease to exist. This second claim sounds a lot like what *I* said about the case.

It would have made more sense for Parfit to claim in this case that it is *indeterminate* whether I cease to exist and *ipso facto* indeterminate whether either of the resulting people would be me. Our concept of a person is just not *precise enough* to have anything to say about cases like this. We can try to *make* it more precise by stipulative improvement. We can try to *force* the concept to respect the principles of logic that lead to the verdict that Parfit claims to be the best description of the case. But, as it stands, our concept leaves open what's true in this case. Perhaps this *really is* what Parfit is saying in this older passage; I just think it could have been put more clearly.

But I do think this is the view Parfit would *now* take about this case. Getting back to the original point about his latest view, he now wants to make this sort of move about many cases, including the case in which a later entity is psychologically perfectly continuous with some earlier person, but where this entity's psychology is transplanted into a new brain. Consider this passage:

> I have discussed cases where Thomson's view seems to me determinately false. In the remaining cases, her view is not, I believe, determinately true. In
>
> > Case (10): My body is destroyed, and a replica created. The resulting person is psychologically continuous with me.
>
> On my view, it would be indeterminate whether the resulting person would be me, or be a new person. But this is one of the cases where, though there is indeterminacy, one description would be most convenient. It would be best, I suggested, to call my Replica a new person (RP: 205). On Thomson's view, that is not merely the best description; it is straightforwardly true….[21]

But *here* I don't see why Parfit clings to the incomplete view that comes from (1) and (2/2*). As he himself admits, "in this case, many people would find Thomson's view more plausible than [his]".[22] But it is not as if we have to adopt Thomson's view to get the right verdict about this case! We could instead accept my Hybrid Theory, which determinately predicts that you cease to exist in this case because *while* there is some psychologically continuous successor, that successor's mental life is not sustained by any of the same matter that sustained your earlier mental life. It seems *obvious* that this is the right thing to say about cases like this.

---

[20] Parfit (1984: 260); italics mine.
[21] Parfit (2008: 184).
[22] Ibid.

2. *Why We Can't Accept a Purely Psychological View*

Enough about the differences between my Hybrid View and Parfit's similar but incomplete view. Let's return to the simpler question of why we ought to prefer a Hybrid View to various *pure* views which say that either psychology or certain purely physical factors are what completely determine numerical personal identity over time. And we'll start by seeing why the purely psychological view must be rejected.

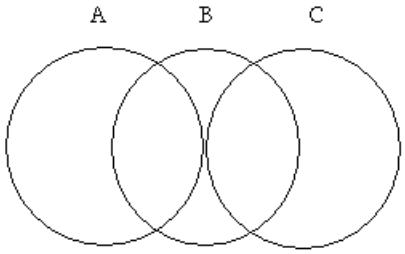
The simplest somewhat plausible version of the pure psychological view is generated by dropping conditions (ii) and (iii) from the Hybrid Theory. It amounts to this claim:

> *Simple Pure Psychological View*: X at *t* is the same person as Y at *t** iff Y is psychologically continuous with X.

When psychological continuity is properly understood, this view does avoid the worries about transitivity that plagued Locke's version of the memory view.

Psychological continuity, remember, is different from psychological connectedness. Connectedness is typically analyzed in terms of *similarity*, so that B is psychologically connected to A to the extent that B is psychologically *like* A. If we understood continuity as mere connectedness, the Simple Pure Psychological View would be crazy. My four year-old self is not psychologically similar to my current self. If continuity were mere connectedness, this would imply that I am a numerically distinct person from my five year-old self, and that that earlier self *literally died*. But it *didn't* die in any more than a metaphorical sense.

More strikingly, if continuity were mere connectedness, the relation of numerical personal identity would cease to be *transitive*.[23] After all, there are cases where A is psychologically similar to B, and B is psychologically similar to C, but where A is not psychologically similar to C. This can happen when the overlap between A and B is completely different from the overlap between B and C, which we can depict in this Venn diagram:




In a case with the structure depicted by this diagram, A would be somewhat similar to B, and B would be somewhat similar to C, but A would not be even slightly similar to C. As it happens, I think this type of pattern can show up in someone's life. My infant self had certain psychological features in common with my five year-old self. And my five year-old self had certain psychological features in common with my ten year-old self. But my infant self and my ten year-old self were barely psychologically similar.

---

[23] A relation **R** is *transitive* when it is true that if (i) A bears **R** to B and (ii) B bears **R** to C, (iii) A bears **R** to C.

We can avoid transitivity failure by distinguishing between continuity and connectedness. Here is the analysis of continuity in terms of connectedness I gave in the last meeting:

> *Definition of Psychological Continuity*: Two individuals A and B existing at distinct times *t* and *t** are psychologically continuous iff there is some series of person-stages $X_1$,…, $X_n$ such that A is strongly psychologically connected to $X_1$, $X_1$ is strongly psychologically connected to $X_2$, $X_2$ is strongly psychologically connected to $X_3$, … and $X_n$ is strongly psychologically connected to B.

This, then, is how we ought to understand continuity. So understood, the Simple Pure Psychological View implies none of the wild conclusions that Locke's memory view implied.

But it still faces problems. The biggest problems surround the possibility of psychological duplication. Suppose that I create a molecule-for-molecule duplicate of all of your body below the head, and then copy the contents of your mind to this new entity's previously *blank* brain. All the while, I leave you intact. Of course, you cannot be identical both to your current self and to this entity. But both are equally psychologically connected with you. Accordingly, the Simple Pure Psychological View implies that you are both. This is not so. So, this view must be revised.

The simplest way of revising it is to add something akin to clause (iii) from the Hybrid Theory. This is a move that Thomson calls adding a "no competitors clause".[24] Now, on *Thomson's* way of revising the Simple Pure Psychological View, the result ends up looking like this (where I vary her formulation slightly to make the result a bit more plausible):

> *Less Simple Pure Psychological View*: X at *t* is the same person as Y at *t** iff Y is psychologically continuous with X, and there is no Z at *t** such that Z ≠ Y and that Z is equally psychologically continuous with X.

Thomson objects that this view is circular. The terms 'X', 'Y' and 'Z' all pick out persons. But if that's right, then the non-identity symbol '≠' would seem to have to pick out the relation of not-being-numerically-personally-identical. But what we are trying to do is to analyze one of the components of this complex negative relation. We cannot appeal to a relation that contains the relation of numerical personal identity as a constituent in an analysis of that very relation.

Parfit gives an answer to this objection, but I have an even simpler answer. In stating a theory of personal identity over time, we *can* appeal to relations of identity between things that are not *already* defined to be persons. If there is a *substantive debate* to be had between Thomson and Parfit, we cannot simply *assume* that the concept of a *mind* or a *mental life* just is the concept of a *person*. Thomson would reject this claim. She would be wrong, but she wouldn't be wrong *by definition*. If that's right, then in stating an analysis of numerical personal identity, we *can* talk of relations of numerical identity between *minds*, since it's supposed to be an *open question* whether people *are* their minds. But if this is right, we can restate the view in the following way that avoids Thomson's objection:

> *Better Less Simple Pure Psychological View*. X at *t* is the same person as Y at *t** iff (i) there are some minds M and M* such that M is X's mind, M* is Y's mind, and M and M* are psychologically continuous, and (ii) there is no mind M** such that M** ≠ M* and that M** is equally psychologically continuous with M.

---

[24] See Thomson (2008: 165) for this.

This version of the view appeals only to the relations of numerical identity and non-identity between *minds* It is not circular to appeal to this relation in giving an analysis of numerical personal identity, since everyone should grant that it's an *open question* whether people are their minds, and whether we can individuate minds independently of the people who have them. If these are open questions, the concept of a mind does not *definitionally presuppose* the concept of a person, and *ipso facto* the concept of numerical identity between minds does not definitionally presuppose the concept of numerical identity between persons. Accordingly, there is no circularity in this revision of the view. This is how I suggested we do things from the beginning. In providing my account of the Hybrid Theory, I appealed to relations of identity between *mental lives*. And all I meant by "mental life" was "mind".

So, anyway, I think a purely psychological theory faces no concerns about circularity. But if not, what *other* concerns might it face? Thomson appears to think that it uniquely faces a concern about *irrelevant extrinsicness*. In discussing a case that could involve the reprogramming of one or two brains to have the same mental life as some guy Brown, she says:

> A second difficulty…is familiar. One who thinks that psychological connectedness without competitors is the mark of personal identity thinks that a question about personal identity can be settled by an appeal to a fact that should surely be irrelevant to it. On this view, neither survivor is Brown if both reprogrammings succeed; but if only one succeeds, then the survivor of that one is Brown. So a small, chance slip-up in the procedures, which fixes that only one reprogramming goes through, fixes that Brown survives. This sounds very implausible.[25]

Parfit rightly notes that this objection overgeneralizes. Thomson's claim that the view makes identity depend on facts that are "surely irrelevant" is itself surely mistaken. For there are other cases in which we definitely want to claim that relations of identity depend on extrinsic factors:

> Properly understood, this conclusion is not, I think, absurd. It will help to consider a simpler pair of cases: those involving Hobbes's famous *Ship of Theseus*. In *Case One*, this ship is dismantled, plank by plank, and is later reconstructed by antiquarians. In *Case Two*, when each plank is removed, it is replaced, so that a working ship continues to exist, though it becomes entirely made of new planks. On what Nozick calls the *Closest Continuer* view, in Case One, the antiquarian's ship *is* the original ship, which has been reconstructed. In Case Two, since the continuously working ship is a closer continuer, it is claimed to be the original ship, and the antiquarian's ship, though made of the same planks, is here claimed to be a different ship.[26]

Here, whether the ship that's reconstructed by the antiquarians is cross-temporally numerically the same as the earlier ship depends on extrinsic facts (i.e., facts not having to do with the *internal features* of the antiquarians' ship), since it may or may not be the Ship of Theseus depending on whether, when Theseus was using it, the original planks were slowly replaced by new ones. There is nothing mysterious or implausible here. We do not want to say that extrinsic factors are "surely irrelevant" to whether the antiquarians' reconstructed ship is the Ship of Theseus. And if extrinsic factors *can* be relevant to questions of identity, Thomson's objection fails.

---

[25] Thomson (2008: 165).
[26] Parfit (2008: 190).

All the same, I *do* think there are clear objections to purely psychological views, and that they straightforwardly motivate moving to the Hybrid View. Indeed, I've already mentioned a case that I think straightforwardly refutes *any* purely psychological view. It is a case in which your entire body, including your brain, is destroyed but is instantly replaced with a duplicate constructed from different physical matter whose brain sustains a mental life that's exactly like your earlier mental life. In this kind of case, you cease to exist. You are replaced by an impostor – an extremely convincing impostor, but an impostor all the same. If you knew that you were going to find yourself in this case, you would regard it as a case in which you would die. Being replaced by a copy is not enough for resurrection. But simple pure psychological views predict otherwise, since the copy will be perfectly psychologically continuous with your earlier self.

The only way we're going to be able to explain this is by appealing to physical factors *somewhere*. Of course, this isn't yet a decisive argument for the Hybrid Theory, since a purely physical theory *would* make the right predictions about this case.

3. *Why We Can't Accept a Purely Physical View*

But purely physical views face different and even more serious objections than purely psychological views. To bring out the problems, let's start with Thomson's

> *Simple Bodily View*, on which X at *t* is the same person as Y at *t** iff X's body is numerically identical to Y's body.

This is an implausible view to which there are many objections. Here's one from Parfit:

> Thomson might…claim that, if my head continued to exist, after the rest of my body was destroyed, this would amount to my body's continuing to exist, though in a diminished state. On this version of Thomson's view…I would still exist.
>
> [But] [o]ur story might continue. Suppose that, after another operation, the blood going to my head came, not from a heart-lung machine, but from someone else's heart and lungs. And suppose that my head was then grafted onto the rest of that other person's body. That other person we can assume to be Thomson, whose head had earlier been destroyed in some accident.
>
> Would these further operations make it true that I would cease to exist? Thomson's answer must be Yes. She believes that, if her body was given a new brain, it would still be the same body, and she would therefore still exist. It could not affect the identity of Thomson's body if its new brain retained its covering of bone and skin. Thomson's view thus implies that, at the end of our story, it would be her who would have my head.[27]

Another objection I find decisive involves cases of complete brain death where someone's body continues to function with minimal external support. Jeff McMahan discusses such a case:

> In one instance, a boy of four was diagnosed as brain dead from intracranial edema caused by meningitis. The physicians recommended discontinuation of life support, but the mother refused. Eventually the boy's body was transferred home where, with only mechanical ventilation, tube feeding, and little more than basic nursing care, it has remained comprehensively functional for the last fourteen years. Alan Shewmon was recently allowed to perform an examination. He reports that

[27] Parfit (2008: 178 – 179)

> "evoked potentials showed no cortical or brain-stem responses, a magnetic resonance angiogram showed no intracranial blood flow, and an MRI scan revealed that the entire brain, including the stem, had been replaced by ghost-like tissues and disorganized proteinaceous fluids." Yet Shewmon also observes that "while 'brain dead' he has grown, overcome infections and healed wounds".[28]

In this case we want to say that the boy's *body* continues to exist. But particularly given the information that *nothing whatsoever* is left of his brain (not even the stem!), do we want to claim that the *person* that once inhabited this body lives? We do not. So people are not their bodies.

You could *try* to claim that the body isn't really living here either, simply because it's being partially externally supported. But this would backfire immediately as an attempt to rescue Thomson's view. Suppose it's true that a body which is supported by this level of external assistance is literally dead. It would then follow from Thomson's view that *even if* there were a perfectly functioning brain in this body (which simply needed some external support), the person who once inhabited the body is dead. There is literally no one there. This is clearly false.

We cannot accept the Simple Bodily View. That doesn't entail that we can't accept a different purely physical view. Many people have been attracted to views on which we are essentially our brains – or, more specifically, our *higher* brains (e.g., the cerebrum and cerebral cortex), which are the parts that most crucially support the capacity for consciousness and intellectual functioning. Since we can understand what a brain is in purely physical terms, we could *ipso facto* understand what persons are in purely physical terms if persons were brains.

Is this theory plausible? Not quite. There is a part of the lower brain known as the *ascending reticular activating system* or *reticular formation* that needs to be intact for the *activation* of consciousness. It is not the part of the brain in which consciousness is *located*: that is the higher-brain. But it must be intact for the higher-brain to sustain consciousness. If we imagined a case where someone's reticular formation was irreversibly damaged, and could not be replaced by a new one without destroying the rest of the person's brain, we would have a case in which the higher brain and most of the lower brain could remain intact, but in which consciousness was irreversibly lost. The same intuitions that motivated moving from a Simple Bodily View to a brain-based view will motivate claiming that the *person* in this case is indeed dead, though virtually all of the former person's body continues to live.

You might try to save the brain-based view by identifying persons with the conjunction of their higher-brain and the reticular formation. On this view, a person lives iff both the higher-brain lives and the reticular formation lives. But in *some* cases, the functioning of the reticular formation can be replaced by external supports, so that it ceases to function but the higher-brain's capacity for consciousness is triggered and active. On the revised view, these would be cases in which the person is dead. But that would be clearly false. This isn't at odds with the earlier case: that was just a case where there was no way to *get* external support to the higher-brain without destroying it. Such cases are imaginable, and refute the simpler view.

You might try to claim that these are cases of mere *technical* irreversible loss of consciousness. You might say that the *capacity* for consciousness in *some sense* persists. You might claim that this harder case is like a case in which we can't turn on a computer without short-circuiting it and frying its motherboard. The computer in some sense still has the capacity to function perfectly: there are just

---

[28] McMahan (2002: 430).

technical problems in getting this capacity to be realized. There is some plausibility to this way of viewing the case, though the sense in which the person would survive in the case at issue would be a remarkably stripped down sense that would have no moral or prudential importance whatever. We would have no obligation to keep the person's body on life support if it were literally technically impossible to try to activate the higher-brain without destroying it. Moreover, if you saw something like this prospect coming in your future, you would surely not say: "Oh, I'm not worried. I'd still continue to exist." You would surely say: "I would be as good as dead at that point, and it would be a total waste of resources to keep my body alive. Give them to people who can be saved."

But there is an even simpler objection to this family of approaches. The *only motivation* for liking a higher-brain-based view was that this view tracked the closest thing to a physical correlate of the capacity for consciousness. It would be *far simpler* to just directly identify the person with the capacity for a certain type of consciousness, and view personal identity over time as continuity of the capacity for consciousness sustained by a series of at least causally connected material bases.

Moreover, the brain-based approaches are simply *chauvinist*. We could easily imagine a world of silicon-based aliens whose capacities for consciousness were sustained by something completely unlike the human brain. Their mental lives could be just as complex as ours. If we cling to a brain-based view, we would have to deny that these entities could be persons, even if their behavior and conscious lives were exactly like ours, and they just happened to be housed in different physical stuff. This seems arbitrary and pointless. But if there is no single *sort* of physical matter that can house the capacity for consciousness, and lots of different sorts of matter could do the job equally well, we're just going to have to abandon the idea of directly identifying persons with particular physical stuffs. We would do better to embrace the Hybrid View and say that persons are essentially *embodied minds*: mental lives, or easily activated capacities for mental life, that are sustained by *some matter or other*.

**References**

McMahan, Jeff. 2002. The Ethics of Killing. Oxford: Oxford University Press.

Parfit, Derek. 1984. Reasons and Persons. Oxford: Oxford University Press.

Parfit, Derek. 2008. "Persons, Bodies and Human Beings" in Hawthorne, J., Sider, T. and Zimmerman, D. (eds.) Contemporary Debates in Metaphysics. Oxford: Blackwell.

Thomson, Judith Jarvis. 2008. "People and their Bodies" in Hawthorne, J., Sider, T. and Zimmerman, D. (eds.) Contemporary Debates in Metaphysics. Oxford: Blackwell.

**MEETING 12 NOTES**

## 1. Taxonomy of Systematic Ethical Theories: Consequentialism vs. Deontology

Broadly speaking, systematic normative ethical theories fall into two categories: **consequentialism** and **deontology**. Since deontological theories are best defined in opposition to consequentialist theories, it is easiest to start with a taxonomy of consequentialist theories.

There are strikingly many versions of consequentialism. One general distinction is between **act** consequentialism and **rule** consequentialism. Act consequentialist theories say that

> (AC) the rightness of some act A that could be performed in some circumstances C turns wholly on the value of the consequences that would result from A-ing in C.

In contrast, rule consequentialist theories say that

> (RC) the rightness of some act A that could be performed in some circumstances C turns entirely on whether A-ing in C is permitted by a set of rules that, when accepted (or followed) by everyone, lead to valuable consequences in the long run.

AC and RC have substantially different implications. Here is one case to bring this out. Perhaps the objective probability that some act A would have good consequences in C-type circumstances is exceedingly low. Still, suppose that someone performs A, and by chance A happens to have really great consequences. An act consequentialist might claim that this act is right, while a rule consequentialist might claim that this act is wrong, since the general policy of allowing A in circumstances like C isn't actually so great, given the objective chances.

Another more important type of case in which AC and RC have different consequences is one with the following contours:

> *Transplant*. I'm in the hospital for some minor operation. You're a doctor. You know that if you killed me now you could use my organs as transplants for five other people in the hospital. You would thereby save these people. And it is certain that these people will otherwise die, since no other organ donors are available.[29]

Most versions of AC will imply that you ought to kill me even without my consent in this case and use my organs to save the five. Doing that would, after all, clearly have the best local consequences: it is better if five people are saved than if one person gets a minor operation while five die. Some find this counterintuitive. RC does not clearly have this implication. For consider the following fact about the case noted by Parfit (forthcoming):

> Suppose we all knew that, whenever we were in hospital, our doctors might secretly kill us so that our organs could be used to save other people's lives. Even if that risk would be very small, this knowledge would make many of us anxious, and would worsen our relation with our doctors. This relation is of great importance, since we often rely on the judgment of our doctors, and their concern for our well-being, and they may be people whom we expect to help us through the ending of our lives…. If all doctors followed this principle in such cases, a few more people's

[29] I take this example from Parfit (forthcoming: 363). It is, however, a very common kind of example used for underscoring the differences between AC and RC.

> lives would be saved. But the saving of these extra lives would be outweighed by these ways in which it would be bad for us and others if, as we all knew, our doctors believed that it could be right to kill us secretly in this way. We can call this the *Anxiety and Mistrust Argument*.[30]

What the Anxiety and Mistrust Argument suggests is that accepting and following a set of principles that would permit the transplant in *Transplant* would in the long run have clearly bad consequences. It would prevent many people from getting help from doctors, since these people would fear what might happen to them by going. Rule consequentialists can appeal to this type of fact to explain our intuition about the moral status of killing me and using my organs as transplants in *Transplant*.

Another general divide is between what are called **actualist** consequentialist theories and **expectabilist** consequentialist theories. A strong version of actualist act consequentialism would claim that it is right for some person P to perform some act A if and only if (this is abbreviated as "iff") P's A-ing *in fact* brings about the best consequences. A strong version of expectabilist act consequentialism would claim that it is right for some person P to perform some act A iff P *expects* that A-ing would bring about the best consequences. These two theories make different predictions, since someone's expectations could be mistaken.

On what basis might one choose between these two theories? Well, one thought that many people have had is that actualist consequentialist theories are going to be *extremely demanding*. The consequences of any given act – i.e., the difference that this act in fact makes to the world – may extend vastly far into the future. To think that, in any case, we could know exactly what the consequences of some act would be is absurd. But notice that if we don't know what the consequences of our acts are in many cases, actualist act consequentialism also implies that we could rarely know whether we are acting rightly. This is a conclusion that many people have found difficult to accept. The same conclusion does not hold for expectabilist versions of consequentialism, since it is *not* difficult for us to know what we *expect* to be the consequences of our acts. We can easily know *that*, since we typically have easy access to our beliefs. So expectabilist versions of consequentialism may appear to be less demanding.

In fact, however, this common claim to an expectabilist advantage is confused, and the reasons why it's confused turn on a very important distinction on which I will frequently rely.

We must generally distinguish between two questions:

> **(Dis)creditability Question**: Under what conditions is an agent *blameworthy* or *praiseworthy* for performing some act?
>
> **(Im)permissibility Question**: Under what conditions is an act permissible (i.e., right) or impermissible (i.e., wrong)?

Everyone must allow that the answers to these questions can come apart. Consider:

> *Nonculpably False Beliefs.* Zane falsely but faultlessly believes that a certain potion he has would kill Jane when it is in fact a cure for all her ills, which include some otherwise terminal ills that will take her life if she doesn't get this potion. He

[30] Parfit (forthcoming: 363).

> recognizes that giving Jane this potion would be prohibited by decisive moral reasons *if* it had the powers he mistakenly takes it to have. But he gives her the potion anyway, and she drinks it. Her life is saved, much to his later chagrin.
>
> *Misleading Clear Evidence*. Jill has strong (but misleading) evidence that flipping the switch would blow up Bill's house and that Bill is inside. She is in a position to know that she has this evidence, and recognizes that killing Bill would be disfavored by decisive moral reasons in the situation that her evidence suggests to obtain. But she flips the switch anyway; it actually dumps a pot of gold in his house.

In these cases, we want to separate the evaluation of the *agent* from the evaluation of his/her *acts*. In the first case, we surely don't want to say, looking at things from a dispassionate and impartial point of view, that there is *nothing good* to be said about Zane's *act*. There are in fact only good things to be said about this act! After all, it *saves Jane's life*. Who could possibly suggest that Jane ought not to be saved in this case? But if we claim that Zane's *act* is *objectively wrong* or *impermissible*, we are committed to this clearly false suggestion: if Zane refrained from giving her the potion, her otherwise terminal ills would take her life. Surely we don't prefer that this happens. And we can say analogous things about the second case.

In saying these things, have we said anything counterintuitive? We haven't. For claims about the permissibility or impermissibility of *acts* do not by themselves imply anything about the creditability or discreditability of the *agents* who perform these acts. We can still say that Zane is a terrible person, because he did what he believed would kill Jane. We can say the same things about Jill in the second case. People can be blameworthy for doing something that is in fact the right thing to do if their beliefs about this thing are false. Indeed, people can also be praiseworthy for doing something that is in fact the wrong thing to do. Consider:

> *Misleading Clear Evidence II*. Dave the highly competent surgeon has exceedingly strong evidence that performing a certain operation would vastly improve Bob's life. Anyone would agree, and no further evidence is available. So, Dave performs the operation on Bob with Bob's consent. In fact, the operation achieves nothing, and indeed makes Bob significantly worse off.

In this case, the agent gets, as we'd say, an "A for effort". But the outcome is no good. We couldn't have advised Dave to perform the operation knowing what actually was going to happen. His *act*, then does not get an A. It gets a bad grade, since it made Bob much worse off.

Accordingly, the (Dis)creditability Question and the (Im)permissibility Question can receive different answers. It is one thing to evaluate an agent, and call him a scumbag or a saint, and another thing to evaluate an act, and call it the right or wrong thing to do. Scumbags can, by accident, do the right thing, and saints, by accident, can do the wrong thing. But once we've seen this distinction, there no longer seems to be a good reason for preferring an *expectabilist* consequentialist theory to an *actualist* consequentialist theory. These theories are about whether acts are objectively right or wrong. To claim that the former theory is unfairly demanding is in a way a category mistake, since it certainly *doesn't* follow that if someone acts wrongly because he fails to have full knowledge of the huge range of consequences of his act, he *ipso facto* is more of a scumbag than he previously was. He needn't be blameworthy at all. It makes more sense, then, to accept some actualist theory in answering the (Im)permissibility Question, and perhaps to accept some expectabilist theory in answering the (Dis)creditability Question. Since, however,

systematic normative ethical theories are always about the former question, this represents no triumph for expectabilism in the sense defined above.

Anyway, let's move to a different contrast. I called both of the earlier versions of consequentialism 'strong' versions. Why? Well, here are two different versions of actualist act consequentialism, the first of which is much stronger:

> **Maximizing Actualist Act Consequentialism**: It is right for P to A iff P's A-ing in fact brings about the *best* consequences.
>
> **Satisficing Actualist Act Consequentialism**: It is right for P to A iff P's A-ing in fact brings about consequences that are *good enough*.

People also often retreat from maximizing to satisficing versions of consequentialism because maximizing consequentialism seems to be an extremely demanding theory. After all, it is a logical consequence of maximizing actualist consequentialism that if P performs an act that brings about consequences that are just ever so slightly less good than some optimal alternative, he acts *wrongly*. Suppose, for instance, that we could numerically measure goodness, and that the optimal act brings about 1,000,000 units of goodness, while the act that P actually performs brings about 999,999 units of goodness. The theory in question entails that P acts wrongly. If this theory were true, it is highly likely that most of us act wrongly all the time. This might look counterintuitive. So, some people suggest that we retreat to satisficing consequentialism. Of course, the big problem with satisficing consequentialism is that there is a burden on its defenders to explain just how much good is good enough. And there is a significant worry that there is no nonarbitrary answer to this question.

But, once again, I think the reasons for preferring satisficing to maximizing consequentialism are here worthless. Claiming that it would be wrong to do the *act* that produces 999,999 units of goodness instead of doing the act that produces 1,000,000 units of goodness doesn't imply anything about whether the *agent* would be terribly blameworthy. Since the difference is so slight, we can hardly blame the agent here. So, in the relevant sense, maximizing consequentialism actually *isn't* a demanding theory. It doesn't automatically issue strong criticisms of agents for failing to do the absolutely best thing.

Now, let's set these issues aside, and focus on maximizing actualist consequentialism for simplicity's sake. There are also many different varieties of maximizing actualist consequentialism. The main factor that distinguishes between these theories is the underlying theory of **goodness** on which they rest. Utilitarians are consequentialists who endorse the following theory of goodness:

> **Utilitarian Account of Goodness**: Some consequence is good only if it actually promotes people's *well-being*.

What exactly is **well-being**? That's a question on which utilitarians disagree quite a bit. There are lots of subspecies of the Utilitarian Account of Goodness. For simplicity, we can focus on:

> **Hedonism**: Well-being consists in a high pleasure-to-pain ratio.
>
> **Preferentialism**: Well-being consists in a high preferences-satisfied-to-preferences-frustrated ratio.

> **Pluralism**: Well-being consists in possessing a high ratio of intrinsic goods (friendship, knowledge, physical health, pleasure, etc.) to intrinsic bads (strife, ignorance, physical unfitness, pain) in one's life.

Just-plain-old utilitarianism is neutral on whether Hedonism, Preferentialism or Pluralism is true.

There are also non-utilitarian versions of consequentialism, but I'll set them aside and turn to a discussion of deontological theories. As I said, deontological theories are usually defined in contrast to consequentialist theories. Most deontologists would, for instance, accept the following clearly non-consequentialist claim:

> **Deontological Dictum** (DD). For at least some acts A and circumstances C, it can be right to perform A in C even if some alternative A* to A has substantially better (actual or expected) consequences in C, and (perhaps) even if the general policy of allowing A*-ing in C-like circumstances would, if adopted at large, lead to substantially better (actual or expected) consequences.

Deontological theories differ on the score of exactly how many types of acts instantiate DD, and on the score of what explains why these acts instantiate DD.

A good way of bringing out the contrast between theories that accept DD and theories that reject DD is by considering the following kind of case:

> *Fat Man on the Bridge*. A trolley car is speeding down the tracks and is just about to pass under a bridge. Farther down the tracks, five people have been tied down by a maniac and will be killed by the trolley if you, who are standing on top of the bridge, don't do something. The only way you could stop the trolley is by hurling something massive in front of it. You're too slender and couldn't do anything by jumping in front of it, and there aren't any big boulders or anything else inanimate that you could throw in front of it either. There is, however, an enormous man standing directly above where the train will pass under. You don't have enough time to persuade him to jump, but you could run and push him off the bridge. It is certain that, if you did this, you would stop the trolley and thereby save the five people further down the tracks.

Pushing the fat man off the bridge would seem to have substantially better consequences than failing to do so in this case, as would one general rule with which this act complies (i.e., *kill one if it would save five*). So, at least on many versions of act consequentialism, it will follow that it would be right to push the fat man off the bridge. This is a result that many find intolerable, and provides a seemingly clear argument for the Deontological Dictum. What explains why we're so willing to reject consequentialist reasoning here? Deontologists point to:

> **Mere Means Principle**. It is never permissible to use another person as a *mere means* to some end, even if the end is very good.

Deontologists often endorse this kind of principle, and claim that it explains our intuitions about *Fat Man on the Bridge*.

In fact, *Fat Man on the Bridge* motivates a stronger claim than WDD:

> **Stronger Deontological Dictum** (DD+): For at least some acts A and circumstances C, it can be impermissible to A even if A has substantially better (actual or expected) consequences in C than every alternative, and even if the general policy of permitting A in C-like circumstances would, if adopted at large, lead to much better (actual or expected) consequences than alternative policies.

This is because most people think that you wouldn't merely be *permitted not* to throw the fat man off the bridge, but that you would *not be permitted* to throw him off the bridge.

Importantly, not all deontologists are willing to slide from DD to DD+, even though DD+ seems necessary to fully explain gut reactions to cases like *Fat Man on the Bridge*.

Indeed, some deontologists are motivated by different considerations than cases like *Fat Man on the Bridge* and principles like the Mere Means Principle. Recall that I said that one complaint about maximizing act consequentialism is that it's overly demanding, and that this is why some people turn to satisficing act consequentialism. It's arguable that plenty of versions of satisficing act consequentialism are still overly demanding. One can bring this out by considering the intuitive thought that if a billionaire fails to donate a very substantial portion of his income to charities like Oxfam, he really isn't even doing *enough* good on impartial grounds: he could save many, many lives and not even have to sacrifice a very decent standard of living by doing this, and so it seems clear that the act of donating is *very* substantially better than the act of holding onto the money and not really doing anything of significance with it. Some deontologists think that the fact that it seems very odd to say that it's *morally impermissible* for the billionaire to fail to donate a very large portion of his income to charities is a further reason for endorsing DD.

This is a different motivation for DD. It is often stated more generally as follows:

> **Agent-Centered Prerogatives**. Sometimes performing the impartially best act or even the impartially good enough act would be very demanding to the agent, and may interfere with the unity and integrity of his life. In these cases, there are *agent-centered prerogatives* that give agents the right to choose to refrain from performing the impartially best or impartially good enough act, even though it may still be permissible to perform that act.

Agent-centered prerogatives are arguably built into certain ordinary normative concepts. Most people believe in such a thing as *supererogation* – i.e., in *going beyond the call of duty*. Mother Theresa, for instance, spent most of her life going beyond the call of duty, and did things that, though amazingly good, were not morally required of her. The very idea of a supererogatory act seems to presuppose the coherence of (weak) agent-centered prerogatives. After all, a supererogatory act is clearly *significantly morally better* than a "merely dutiful" alternative; such an alternative may clearly not only fail to be the best act, but fail to get even close to being the best act.

The attempt to motivate DD by the existence of agent-centered prerogatives is to be contrasted with the earlier consideration, which is stated more generally as follows:

> **Agent-Centered Restrictions**. Sometimes performing the impartially best act or even the impartially good enough act would require one to do certain intuitively objectionable things to some subclass of people – e.g., to use the fat man as a mere means to the impartially optimal end of saving five lives. In these cases, there are

> *agent-centered restrictions* that prevent agents from performing the impartially best or impartially good enough act, and that make it impermissible to perform that act.

As I said in less technical language before, agent-centered restrictions seem to be needed to make full sense of *Fat Man on the Bridge*.[31]

Even so, some deontologists are willing to follow act consequentialists in rejecting agent-centered restrictions while departing from them in accepting prerogatives. This position is neither obviously incoherent nor ill-motivated. Why? Well, although they may seem to be needed to capture some gut reactions, restrictions give rise to seeming paradoxes to which prerogatives do not. A full-fledged defender of restrictions will claim that it's wrong to murder one not just to prevent several other *deaths* or *lettings-die* from occurring (as in *Fat Man on the Bridge*), but also to prevent several other *murders* from occurring.

This is a peculiar claim. What is it about the fact that *I* am the murderer of the one that makes the state of affairs in which I murder one and no one else murders *objectively morally worse* than the state of affairs in which I murder none and someone else murders four? Some philosophers – e.g., Samuel Scheffler – think that there is no good answer to this question, but still feel compelled by considerations of personal integrity and demandingness to accept prerogatives.

In any case, the upshot is that we can distinguish three versions of deontological ethics:

> **Weakest Deontology**: There are strong agent-centered prerogatives (e.g., permissions to do acts that aren't even good enough by impartial lights) but no agent-centered restrictions.
>
> **Middling Deontology**. There are agent-centered restrictions but no strong agent-centered prerogatives.
>
> **Strongest Deontology**. There are agent-centered restrictions and strong agent-centered prerogatives.

Kant and his followers generally accept either Middling or Strongest Deontology. But there are some – e.g., Scheffler – who have toyed with Weakest Deontology. Next week I'll talk a lot more about Kant's version of deontological ethics and the specific debates that arise between Kantian deontologists and act consequentialists like Mill.

## 2. Agent-Centered Restrictions and the Creditability/Permissibility Distinction

For the moment, it is worth talking at a higher level of generality about what motivates agent-centered restrictions.

As I've already noted, these restrictions can seem very puzzling. In *Fat Man on the Bridge*, more lives would be saved if you threw the fat man off the bridge than if you refrained from doing so. Deontologists say – reflecting, I take it, common sense morality – that this act is still impermissible. But what makes it wrong to throw the fat man off here? It can't just be the disvalue that attaches to his *death*, since that disvalue is clearly outweighed by the fact that *four more* people will die otherwise. It also cannot just be the disvalue that attaches to his being

[31] The terms "agent-centered prerogative" and "agent-centered restriction" were introduced by Scheffler (1982).

*murdered*, since, in a variation of the case, we could imagine that some maniac deliberately started the trolley in motion with the intention to kill the five; if you failed to intervene here, he would successfully murder them, and so a greater number of murders would occur. What could the difference be? Deontologists here often just appeal to more distinctions in commonsense morality – distinctions, I believe, that need as much justification as the particular agent-centered restrictions that they are invoked to motivate.

One general distinction that deontologists think is morally significant is the distinction between *doing* and *allowing*. And it may seem to be a feature of our commonsense moral thought that it is less permissible to *do* bad than to *allow* bad to be done. We think that allowing a starving person to die is clearly not as impermissible as actively starving someone. If this were a bedrock intuition, we would be forced to accept particular agent-centered restrictions like the one that commonsense supports in the case of *Fat Man on the Bridge*. After all, in this case, even if the outcome would be worse if the five were killed, at least *you yourself* wouldn't be *actively causing* it.

But now it is worth reflecting on a distinction we made earlier – viz., between act-oriented features like *impermissibility* and *permissibility* on the one hand, and agent-oriented features like *blameworthiness* and *praiseworthiness* on the other. As I noted, we have to allow that these features can come apart, so that someone can act impermissibly but not be blameworthy for it, and act permissibly but not be praiseworthy for it. Once this point is appreciated, it becomes much harder to see why we ought to follow the deontologist in claiming that the distinction between doing and allowing tracks some difference in *permissibility* rather than simply in *blameworthiness*. We can grant that actively causing harm is often more blameworthy than passively allowing it. This is, however, compatible with allowing that both are just as impermissible.

We can be pressured into accepting this view about doing vs. allowing. For in *some* cases, this distinction tracks no difference in permissibility, and it seems like the only disanalogy between this case and other cases is that the agent lacks some *excuses* – i.e., some *blame-relieving* appeals – that he doesn't always have. Perhaps the best example of this was constructed by James Rachels in the course of his argument that the distinction between killing and letting die cannot generally track a difference in permissibility.

Rachels pointed to the following cases:

> *Smith's Case*. Smith stands to gain a large inheritance if anything should happen to his six-year-old cousin. One evening while the child is taking his bath, Smith sneaks into the bathroom and drowns the child, and then arranges things so that it will look like an accident. No one is the wiser, and Smith gets his inheritance.
>
> *Jones's Case*. Jones also stands to gain if anything should happen to his six-year-old cousin. Like Smith, Jones sneaks in planning to drown the child in his bath. However, just as he enters the bathroom, he sees the child slip, hit his head, and fall face-down in the water. Jones is delighted; he stands by, ready to push the child's head back under if necessary, but it is not necessary. With only a little thrashing about, the child drowns all by himself, 'accidentally', as Jones watches and does nothing. No one is the wiser, and Jones gets his inheritance.[32]

[32] These cases are taken *verbatim* from Rachels (1986: 112).

As Rachels notes, if we really think that the distinction between killing and letting die has fully general significance with respect to questions of *permissibility*, we ought to believe that Smith acted *less permissibly* than Jones. But we don't think this. So, the distinction cannot have this general significance. And the reason why it doesn't in this case is clear. What is unusual about Rachels' cases is that both Smith and Jones are terrible agents with blameworthy intentions; usually, an agent in a paradigm case of killing will have more blameworthy intentions than an agent in a paradigm case of letting die. Rachels's cases helpfully prevent us from conflating impermissibility and blameworthiness because there is no difference in blameworthiness. And, as predicted, if we fix this confounding factor, we aren't inclined to think that there should be any further difference in permissibility.

I think this type of point can be used to systematically undermine other attempts to support agent-centered restrictions by appeal to certain distinctions in commonsense morality. A distinction that is much like the distinction between doing and allowing harm is the distinction between *intending* and *merely foreseeing* bad effects.

One can try to bring out the intuitive force of this distinction by considering the difference between terror bombing and tactical bombing. If one pilot drops bombs on some area with the intention of killing innocent civilians and also happens to destroy some genuinely bad military base just by chance, we would regard this pilot as a terrorist or war criminal. If, on the other hand, some pilot drops bombs on an unjust opponent's base, but foresees that he will also kill some innocent civilians as a byproduct of this and deeply regrets this fact, we might regard him as an honorable strategist who chose the lesser of two evils. In this type of case, the only difference between the two seems to be that the former intends the bad states of affairs, whereas the latter merely foresees them as an unfortunate byproduct of some ultimately just cause.

And this can seem morally significant. But there's a strong argument that the moral significance cannot be at the level of permissibility. Judy Thomson does the best job of bringing this out:

> Suppose a pilot comes to us with a request for advice: "See, we're at war with a villainous country called Bad, and my superiors have ordered me to drop some bombs at Placetown in Bad. Now there's a munitions factory at Placetown, but there's a children's hospital there too. Is it permissible for me to drop the bombs?" And suppose that we made the following reply: "Well, it all depends on what your intentions would be in dropping the bombs. If you would be intending to destroy the munitions factory and thereby win the war, merely foreseeing, though not intending, the deaths of the children, then yes, you may drop the bombs. On the other hand, if you would be intending to destroy the children and thereby terrorize the Bads and thereby win the war, merely foreseeing, though not intending, the destruction of the munitions factory, then no, you may not drop the bombs." What a queer performance this would be![33]

As Thomson usefully suggests here, questions of permissibility are questions about what to do. Your intentions are independent of what you're going to do, and it is clearly strange to think that, in deciding whether some object of intention would be impermissible, you would *look inward* at your own motivations. Of course, this isn't to say that they are entirely morally irrelevant: they are simply relevant to a *different question* – namely, the question of your *blameworthiness*. Once we refine our intuitions to track the difference between

[33] Thomson (1991: 293).

impermissibility/permissibility and blameworthiness/praiseworthiness, we're going to have a hard time stomaching the sorts of agent-centered restrictions to which Strongest and Middling Deontology appeal. We'll see more of this when we turn to Kant briefly next week.

**References**

Parfit, Derek. Forthcoming. On What Matters. Oxford: Oxford University Press.

Scheffler, Samuel. 1982. The Rejection of Consequentialism. Oxford: Oxford University Press.

Thomson, Judith Jarvis. 1991. "Self-Defense." *Philosophy and Public Affairs* 20.4: 283-310.

## 1. Why Agent-Centered Restrictions (and Some Deontological Views) are Dubious

### 1.1. *The Unreliability of Our Deontological Gut Reactions and a Psychological Debunking Explanation*

As we've seen, a key difference between all consequentialist theories and those deontological theories that seem most in sync with commonsense moral intuition is that the latter views embrace this claim:

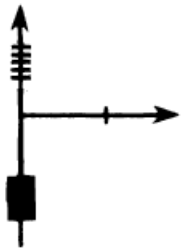
> **Agent-Centered Restrictions**: Sometimes performing the impartially best act or even the impartially good enough act would require one to do intuitively objectionable things to some subclass of people. In these cases, there are agent-centered restrictions that make it impermissible for agents to perform the impartially best or impartially good enough act.

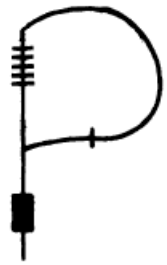
Positing agent-centered restrictions offers a direct way of explaining gut feelings about cases like:

> *Fat Man on the Bridge*. A trolley car is speeding down the tracks and is just about to pass under a bridge. Farther down the tracks, five people have been tied down by a maniac. They will all be killed by the trolley if you, who are standing on top of the bridge, don't do something. The only way you could stop the trolley is by hurling something massive in front of it. You're too slender and couldn't do anything by jumping in front of it, and there aren't any big inanimate things that you could throw in front of it either. But there is an enormous man standing directly above where the train will pass under. You don't have enough time to persuade him to jump, but you could run and push him off. It is certain that, if you did this, you would stop the trolley and thereby save the five people farther down the tracks.

In one clear sense, the best outcome would be if the five were saved: it is better *from an impartial point of view* if five live and one dies than if one lives and five die. But before acquainting ourselves with the arguments of some consequentialists, it is extremely tempting to claim that pushing the fat man off in this case is still simply wrong regardless of its impartial consequences.

Still, this intuitive reaction is hard to justify, particularly when we compare *Fat Man on the Bridge* with other cases. Consider the following two cases, which differ only in a single detail:

| | |
|---|---|
| *Bystander at the Switch*. A trolley car is speeding down the tracks and is approaching a junction. If the car stays on the same track, it will run over five slender people who have been tied down. But if it is diverted to the other track at the junction, it will only kill an enormous man who has caught himself on the track and can't get off. You are a bystander near the junction who is aware of all these facts. You are also in a position to flip the switch that would send the train down the other track. |  <br>From Thomson (1985: 1402) |

| | |
|---|---|
| *Loop*. Everything is just like in *Bystander at the Switch* except for this detail: the second track at the junction eventually loops back onto the first track just before the bit where the five are tied down. For this reason, if the fat man weren't caught in the middle of the second track, you would not be able to save the five. You are in a position to save them only in virtue of his presence, and you are aware of this fact. |  <br>From Thomson (1985: 1402) |

Most of us have the intuition in *Bystander at the Switch* and *Loop* that it would be permissible to flip the switch. Still, if you flip the switch, you are killing the one. So we cannot explain the gut reaction about *Fat Man on the Bridge* by appeal to the distinction between *killing* and *letting die*, and say that while you'd only be letting the five die, you would be positively killing the one. For you also kill in *Bystander at the Switch*, albeit not exactly "with your bare hands"; but surely *that* latter detail is irrelevant, since shooting someone can be wrong but not involve any bodily contact.

How else could we explain the difference we seem to intuit between *Bystander at the Switch* and *Fat Man on the Bridge*? I noted before that a common thing that deontologists in the past cited as being amiss with *Fat Man on the Bridge* is that you are killing the fat man as a *means* to the optimal end of saving the five. They then say that this principle justifies our intuition about that case:

> **Mere Means Principle**. It is never permissible to use another person as a mere necessary means to some end, *even if* the end is impartially optimal or nearly so.

Ironically, however, you are *also* using the fat man as a necessary means for an impartially optimal end in *Loop*. You take advantage of the fat man's presence on the second looping track as a method for saving the five in *Loop*. After all, if he weren't there, you couldn't save the five. What remains abundantly clear is that we do think you are permitted to flip the switch in *Bystander at the Switch*. Since the only difference between that case and *Loop* is the presence of the extra bit of track, and that doesn't seem relevant (a fact that is reflected in our intuitions, which also suggest that switching in *Loop* is permissible), we have to reject the application of the Mere Means Principle in this case if we want to capture the intuition. Yet the Mere Means Principle seemed like the only thing that could explain our gut reaction in *Fat Man on the Bridge*.

This suggests that the gut reaction wasn't reliable. Indeed, a *debunking explanation* of that reaction is available which reveals why it should be instilled in us for reasons having nothing to do with the moral facts. Empirical research suggests that the main factor that explains why it is that people have the strong judgment of permissibility in *Fat Man on the Bridge* but lack it in *Loop* is that the first involves imagining a very *direct kind of engagement* with the person, while the latter does not. These facts themselves must be irrelevant, since it is just as bad to press a button to deliberately blow up a distant man as it is to beat him to death with one's bare hands. Peter Singer summarizes some of the implications of data discovered (inter alia) by Princeton psychologist Joshua Greene in an article that calls some commonsense reactions into question on empirical grounds:

> Let us return for a moment to the trolley problem cases. As mentioned before, philosophical discussions of these cases from Thomson onwards have been preoccupied with the search for differences between the cases that justify our initial intuitive responses. If, however, Greene is right to suggest that our intuitive responses are due to differences in the emotional pull of situations that involve bringing about someone's death in a close-up, personal way, and bringing about the same person's death in a way that is at a distance, and less personal, *why should we believe that there is anything that justifies these responses*? If Greene's initial results are confirmed by subsequent research, we may ultimately conclude that he has not only explained, but *explained away*, the philosophical puzzle….
>
> This becomes clearer when we consider how well Greene's findings fit into the broader evolutionary view of the origins of morality…. For most of our evolutionary history, human beings have lived in small groups, and the same is almost certainly true of our pre-human primate and social mammal ancestors. In these groups, violence could only be inflicted in an up-close and personal way – by hitting, pushing, strangling, or using a stick or stone or club. To deal with such situations, we have developed immediate, emotionally based responses to questions involving close, personal interactions with others. The thought of pushing the stranger off the footbridge elicits these emotionally based responses. Throwing a switch that diverts a train that will hit someone bears no resemblance to anything likely to have happened in the circumstances in which we and our ancestors lived. Hence the thought of doing it does not elicit the same emotional response as pushing someone off a bridge. So the salient feature that explains our different intuitive judgments concerning the two cases is that the [bridge] case is the kind of situation that was likely to arise during the eons of times over which we were evolving; whereas the standard trolley case describes a way of bringing about someone's death that has only been possible in the past century of two, a time far too short to have any impact on our inherited patterns of emotional response.[34]

For these reasons, we should take reliance on the gut reaction about *Fat Man on the Bridge* to be controversial at best. For all the clearly ethically significant features of the case are preserved in *Loop*, and yet we *don't* think that there is anything impermissible about flipping the switch in *Loop* – and reasonably so, since it is not *relevantly* different from *Bystander*, and it is obviously permissible to flip the switch in *Bystander*. This kind of reasoning casts serious doubt on the foundations of deontological ethical theories.

### 1.2. *A Philosophical Debunking Explanation of Our Deontological Gut Reactions*

There are further problems that I didn't get to last time. They center on the fact that it looks like the key notions deontologists use to try to vindicate the claim that there are some agent-centered restrictions conflate *impermissibility* with *blameworthiness*, properties that are clearly distinct.

Note that one general distinction that deontologists think is morally significant is the distinction between *doing* and *allowing*. It may seem to be a feature of our commonsense moral thought that it is less permissible to *do* bad than to *allow* bad to be done. We think that allowing a starving person to die is clearly not as impermissible as actively starving someone. If this were a bedrock intuition, we would be forced to accept particular agent-centered restrictions like the one that commonsense supports in the case of *Fat Man on the Bridge*. After all, in this case, even if the outcome would be worse if the five were killed, at least *you yourself* wouldn't be *actively causing it*. (But as we saw in *Bystander at the Switch*, this is not reflected in our intuitions about some cases!)

---

[34] Singer (2005: 347-8), italics mine.

But now it is worth reflecting on a distinction I made last time – viz., between act-oriented features like *impermissibility* and *permissibility* on the one hand, and agent-oriented features like *blameworthiness* and *praiseworthiness* on the other. As I noted then, we must allow that these features can come apart, so that one can act impermissibly but not be blameworthy for it, and act permissibly but not be praiseworthy for it. Once this point is appreciated, it becomes much harder to see why we ought to follow the deontologist in claiming that the distinction between doing and allowing tracks some difference in *permissibility* rather than simply in *blameworthiness*. We can grant that actively causing harm is often far more blameworthy than passively allowing it. This is, however, compatible with allowing that both are just as impermissible.

Indeed, we can be very strongly pressured into accepting this view about the significance of doing vs. allowing. For in *some* cases, this distinction tracks no difference in permissibility, and it seems like the only disanalogy between this case and other cases is that the agent lacks some *excuses* – i.e., some *blame-relieving* appeals – that he doesn't always have. Perhaps the best example of this was constructed by James Rachels in the course of his argument that the distinction between killing and letting die cannot generally track a difference in permissibility.

Rachels pointed to the following cases:

> *Smith's Case.* Smith stands to gain a large inheritance if anything should happen to his six-year-old cousin. One evening while the child is taking his bath, Smith sneaks into the bathroom and drowns the child, and then arranges things so that it will look like an accident. No one is the wiser, and Smith gets his inheritance.
>
> *Jones's Case.* Jones also stands to gain if anything should happen to his six-year-old cousin. Like Smith, Jones sneaks in planning to drown the child in his bath. However, just as he enters the bathroom, he sees the child slip, hit his head, and fall face-down in the water. Jones is delighted; he stands by, ready to push the child's head back under if necessary, but it is not necessary. With only a little thrashing about, the child drowns all by himself, 'accidentally', as Jones watches and does nothing. No one is the wiser, and Jones gets his inheritance.[35]

As Rachels noted, if we really think that the distinction between killing and letting die has fully general significance with respect to questions of *permissibility*, we ought to believe that Smith acted *less permissibly* than Jones. But we don't think this. So, the distinction cannot have general significance with respect to permissibility. And the reason why it doesn't in this case is clear. What is unusual about Rachels' cases is that both Smith and Jones are terrible agents with blameworthy intentions; usually, an agent in a paradigm case of killing will have more blameworthy intentions than an agent in a paradigm case of letting die. Rachels's cases helpfully prevent us from conflating impermissibility and blameworthiness because there is no difference in blameworthiness. As my view predicts, if we fix this confounding factor, we aren't inclined to say that there's a further difference in permissibility.

This type of point can be used to systematically undermine other attempts to support agent-centered restrictions by appeal to distinctions in commonsense morality. A distinction that is much like the distinction between doing and allowing harm is the distinction between *intending* and *merely foreseeing* bad effects. One can try to bring out the gut intuitive force of this distinction by considering the difference between terror bombing and tactical bombing. If one pilot drops

---

[35] These cases are taken *verbatim* from Rachels (1986: 112).

bombs on some area with the intention of killing innocent civilians and also happens to destroy a genuinely bad military base just by chance, we would regard this pilot as a terrorist or war criminal. But if a pilot drops bombs on an unjust opponent's base, but foresees that he will also kill some innocent civilians as a byproduct of this and deeply regrets this fact, we might regard him as an honorable strategist who chose the *lesser of two evils*. In this type of case, the only difference between the two seems to be that the former intends the bad states of affairs, whereas the latter merely foresees them as an unfortunate byproduct of some ultimately just cause.

So, this distinction can seem morally significant. But there's a strong argument that the moral significance cannot be at the level of *permissibility*. Thomson did the best job of bringing this out:

> Suppose a pilot comes to us with a request for advice: "See, we're at war with a villainous country called Bad, and my superiors have ordered me to drop some bombs at Placetown in Bad. Now there's a munitions factory at Placetown, but there's a children's hospital there too. Is it permissible for me to drop the bombs?" And suppose that we made the following reply: "Well, it all depends on what your intentions would be in dropping the bombs. If you would be intending to destroy the munitions factory and thereby win the war, merely foreseeing, though not intending, the deaths of the children, then yes, you may drop the bombs. On the other hand, if you would be intending to destroy the children and thereby terrorize the Bads and thereby win the war, merely foreseeing, though not intending, the destruction of the munitions factory, then no, you may not drop the bombs." What a queer performance this would be![36]

As Thomson rightly suggests, questions of permissibility are questions about *what to do*. But the character of your motives is independent of what you are going to do: a scumbag and a saint can intend the same act for different reasons. It is incredibly strange to think that, in deciding whether some object of intention would be impermissible, you should be *looking inward* at your own motivational states. Of course, this isn't to say that they are entirely morally irrelevant: they are just relevant to a *different question* – namely, the question of your *blameworthiness*. Once we refine our intuitions to track the difference between impermissibility/permissibility and blameworthiness/praiseworthiness, we're going to have a hard time stomaching the sorts of agent-centered restrictions on the promotion of optimal ends to which strong deontologists like to appeal. This, together with the data about the trolley cases, suggests that we shouldn't leap to embrace these restrictions on purely pretheoretical intuitive grounds. Our pretheoretical intuitions are clearly unreliable and easily confused.

## 2. Kant's Categorical Imperative and its Problems

Let's turn briefly to a fragment of Kant's ethics that Marcello and James discussed. Typically Kant is classified as a deontologist who accepts strong agent-centered restrictions, but there is an enormous amount of scholarly debate about whether that really is a correct description of his view. I'll give you the traditional "textbook" presentation of Kant, but if you're interested, I suggest reading Parfit (forthcoming: chs.12-17), who gets deep into the subtleties of Kantianism.

One of Kant's formulations of his supreme principle of morality is

> *The Universal Law Categorical Imperative* (ULCI): You ought only to act on those maxims that you could consistently will to be universal laws.

---

[36] Thomson (1991: 293).

Let's unpack the elements of ULCI. By "maxim", Kant means something close to what we would more normally pick out with the term "motive". This should be familiar enough. People can do the same thing for different reasons. One student might choose to take a class simply to get a grade, care little about the content of the class, and so be bound to forget it quickly thereafter. Another student might choose to take the same class because she believes that intellectual flourishing is an essential component of a truly good life, and for this wants reason absorb all the insights she expects to see in the class and take them with her as a guide to later life. Both students make the same choice, but with profoundly different motivations. In some sense, we regard the choice of the latter student as *worthier* than the choice of the former student. But this has nothing directly to do with the *object of choice*, which is the same in both cases – viz., whether to take the class. It has more to do with the motives behind the choice, whose (dis)value seems to indirectly enrich (or diminish) that of the choice. In building the idea of a *maxim* into ULCI, Kant's starting point is this observation about the fact that our judgments of *worth* tend to depend on the motives for which the acts are performed.

Now, what did Kant mean by "will to be a universal law" in ULCI? There are many ways of understanding this phrase, but the one on which I'll focus is this:

> you would will some maxim to be a universal law if and only if you would make it the case that *everyone acts according to the rule to which this maxim gives expression.*

When the crucial phrase is understood in this way, what the Universal Law Categorical Imperative asks us to do is to imagine ourselves as *lawmakers*: we take the motives that guide our acts and imagine turning the rules to which those motives give expression into rules that could govern everyone. Roughly speaking, if the rule to which the motive of our action gives expression could be *consistently universalized* in this way (i.e., be universalized without any *contradiction*), Kant thinks that your act would be permissible. If the rule to which the motive of your action gives expression couldn't be consistently universalized in this way, Kant thinks that your act would be impermissible.

At this point, it is helpful to consider examples to bring out the last element of the Universal Law Categorical Imperative. One of Kant's best illustrations involves a case where someone makes a lying promise. He imagines that some guy intends to get a loan and promises to repay it, but with no real intention of repaying it. Here the maxim of his act is: break your promises whenever it would benefit you. Kant points out that if we turned this maxim into a universal law, so that everyone would break their promises when it would benefit them, and we all knew that this was so, *the practice of promising would cease to exist*: we wouldn't be able to take each other's promises at all seriously if it were so common and easy for us to break them. So, we couldn't really *consistently* will this maxim to be a universal law, because if we were all knowingly disposed to break our promises whenever it would benefit us, we wouldn't have a practice of promise-making at all. The universalized maxim is self-defeating, and in this way *contradictory* in an *informal* sense. If we apply the ULCI to this case, we get the verdict that making a lying promise is wrong. And that is plausible.

Alas, the devil is in the details. There are a great many maxims that can be consistently universalized, but acting on which is clearly wrong. Take some genocidal maxim of the form: kill innocent people in group G. As long as there are sufficiently many non-G people in the world, this maxim could be universalized without any *contradiction*. Still, acting on this maxim is obviously wrong. A different problem for Kant is that *some* maxims cannot be consistently universalized, but acting on them is not at all wrong. Consider the maxim: eat food without

replacing it with newly created food. If we all acted on this maxim, a contradiction in the same informal sense that arose in the case of the lying promise would ensue, because after a pretty short period there would be no more food to eat. Still, it is not morally wrong for some of us not to be food producers. Here it's worth remembering that the driving idea behind Kant's Universal Law Categorical Imperative is nicely expressed in the thought "What if everyone did that?" Sometimes we can use this question to object to someone's act. It works well in the case of the lying promise. Still, it isn't always a good objection, because it's often enough to simply reply: "Some people won't act on this maxim". We only need *some* food producers!

The simple formulation of the Universal Law Categorical Imperative thus fails. Kant did have other principles that he thought were candidates for being supreme principles of morality:

> *Kingdom of Ends Categorical Imperative*: You ought always treat other persons as ends in themselves, and never merely as means.
>
> *Consent Principle*: You ought to treat other people only in ways to which they would have sufficient reasons to consent.

The Kingdom of Ends Categorical Imperative has some plausibility, but its force is undercut by our earlier reflections on cases like *Loop*, and by the fact that *whether* you treat someone as a means depends on your own intentions and beliefs, which seem more relevant to agent-oriented questions of *blame* and *praise* than to act-oriented questions about *permissibility* and *impermissibility*. Indeed, there are some acts that this imperative does not condemn that are clearly wrong. Suppose you have some crazy beliefs: you think that it would be best for some person if that person were dead. You might kill him out of the belief that you are helping him (e.g., by engaging in "euthanasia"). Internally speaking, you are treating him as an end and not as a means: you think you're acting out of concern for his interests for their own sake, and that it just happens to be in his interest to die. The Kingdom of Ends Categorical Imperative fails to condemn this act. But it is clearly wrong all the same.

The Consent Principle is more plausible, but it, too, faces some problems. Go back to *Loop*. In this case, would the fat man on the other track have sufficient reasons to consent to your flipping the switch? This seems at best debatable: his own interests give him strong reasons to object, and he might not unreasonably complain about what you are about to do. Still, it remains intuitively permissible to flip the switch in this case to save the five. The Consent Principle cannot capture this thought without an added theory according to which the reasons that a person possesses to consent to some act are *purely impartial*. That theory would not be obviously right. Without it, the Consent Principle would wrongly condemn flipping the switch in *Loop*.

## 3. Arguments for the Permissibility of Early and Late Abortion

### 3.1. *The Main Argument for the Permissibility of Early Abortion*

To bring out why *early* abortions – i.e., ones before 20-25 weeks, before the fetus has a brain that can support the capacity for consciousness – are permissible, let's consider what's wrong with the following naïve argument for thinking that all abortions, early and late, are impermissible:

> *The Naïve Argument for Impermissibility*
>
> 1. Every innocent living being has the right to life. (Assumption)
> 2. Every fetus is an innocent living being. (Assumption)
> 3. So, every fetus has the right to life. (Follows from 1 & 2)

4. If X has the right to life, then causing X to die violates X's rights. (Assumption)
5. So, causing a fetus to die by aborting it violates its rights. (Follows from 3 & 4)
6. Violating an innocent person's rights is always impermissible. (Assumption)

---

7. So, abortion is always impermissible. (Follows from 5 & 6)

This argument faces a dilemma which I'll first sketch in broad strokes and then unpack in more detail. If "living being" is understood *so inclusively* as to include beings that *lack the capacity for consciousness*, then the argument overgeneralizes to yield clearly unacceptable conclusions (e.g., that killing plants is wrong), and (1) is false. If, on the other hand, "living being" is understood *less inclusively*, so that it includes only beings with the capacity for consciousness, then (2) will not be true of fetuses before 20-25 weeks, when they lack developed brains that could support the capacity for consciousness. Either way the argument would fail. This, in turn, suggests a positive case for the permissibility of early abortions: since early fetuses have no capacity for consciousness, there is *no person* present whose rights we could possibly violate by ending its merely biological existence. If there is no person there whose rights we could violate, we couldn't be acting wrongly.

To bring this out in more detail, let's note that there is an intuitive distinction between *biographical life*, which is the kind of life lived by a conscious or sentient individual, and *biological life* which is the kind of life lived by any organism, including plants and bacteria. We care a lot more about *biographical* life than *biological* life. Just recall a case from McMahan that I brought up in discussing personal identity:

> In one instance, a boy of four was diagnosed as brain dead from intracranial edema caused by meningitis. The physicians recommended discontinuation of life support, but the mother refused. Eventually the boy's body was transferred home where, with only mechanical ventilation, tube feeding, and little more than basic nursing care, it has remained comprehensively functional for the last fourteen years. Alan Shewmon was recently allowed to perform an examination. He reports that "evoked potentials showed no cortical or brain-stem responses, a magnetic resonance angiogram showed no intracranial blood flow, and an MRI scan revealed that the entire brain, including the stem, had been replaced by ghost-like tissues and disorganized proteinaceous fluids." Yet Shewmon also observes that "while 'brain dead' he has grown, overcome infections and healed wounds".[37]

In this case we want to say that the boy's *body* continues to exist. But given the information that *nothing whatsoever* is left of his brain (not even the stem!), do we want to claim that the *person* that once inhabited this body lives? No. If you knew that you would end up in the condition that this boy was going to end up in, what would you say? You'd say: "I'd be as good as dead at that point". If you knew that your later body could be taken off life support, and the medical resources could be used to save other still conscious beings, you would surely, if you're at all morally decent, permit your mere organism to be taken off "life support" to help others. Indeed, many of us would believe that even if you had never known that this was going to happen to you, once it does happen and you have no brain whatsoever left, it would be permissible to redirect the medical resources to sentient beings that could be saved with their mental lives intact. These considerations strongly suggest that we don't think that the kind of

[37] McMahan (2002: 430).

life lived *merely* by the organism we inhabit has any intrinsic moral value. What we think is morally valuable is the conscious being that inhabits the body.

But now reflect on the fact that fetuses before 20-25 weeks have no capacity for consciousness. If we apply the same type of reasoning that led us to conclude that it is permissible (indeed, perhaps *obligatory*, if the number of conscious beings that could be saved is large) to terminate *merely biological* life support in cases like the one discussed by McMahan, we will be led to the conclusion that there could be nothing wrong with an early abortion. For there is no *person* there at all in the early stages where the fetus lacks a developed brain. There is just a purely physical organism with the moral status of the wholly brain dead patient in McMahan's case. Accordingly, if premise (1) in the Naïve Argument for Impermissibility were read so that "living being" meant "biologically or biographically living being", it would have to be false. If, on the other hand, "living being" throughout the argument meant only "biographically living being", premise (2) would be false: fetuses before 20-25 weeks lack the capacity for consciousness and sustain no biographical life.

How might someone respond to this argument? The most common response I've heard appeals to the thought that even an early fetus has the *potential* to become a conscious being. This, however, is not a good reply. Consider the product of a sperm-egg pair just a few hours after conception. What we have is a *zygote*. Most of us do not think that it is impermissible to kill zygotes. After all, many women use the "morning-after" pill. What this pill does is kill the zygote before it develops further into an embryo. Yet the zygote, like the early fetus, has the potential to become a conscious being in exactly the same sense of "potential". The differences between it and an early fetus are *entirely superficial*: the early fetus "looks" more like a human person. (But *looks* don't call the moral shots. In the case cited by McMahan, the wholly brain dead being *looked* like a human person. But that didn't make it impermissible to remove life support and apply the medical resources elsewhere.) More generally, there is a sense in which a sperm-egg pair has the potential to become a conscious being by fusing together and forming a zygote. Yet contraception is surely not morally impermissible. So, the appeal to potential fails. Since it appears to be the only answer to the dilemma just constructed, we're in a position to conclude that early abortions are perfectly morally permissible.

### 3.2. *Thomson on Late Abortion*

Let's turn to Thomson, who has a different focus. She wants to argue that even some late abortions are permissible. Accordingly, she focuses on cases in which a fetus *might* plausibly be claimed to have a right to life, and then attacks (4) and (5) in the *Naïve Argument for Impermissibility* with a series of arguments from analogy. Her first argument turns on:

> *Violinist I.* You've been kidnapped and connected to an unconscious famous violinist, who will die unless he uses your kidneys for the next nine months.

As she suggests, it would be absurd if the director of the hospital said to you: "We're sorry that the Society of Music Lovers did this to you, but we cannot unplug you. After all, this violinist is a person with the right to life, and to unplug him from you would cause him to die, and that would violate his right to life, which is impermissible." But the director of the hospital would be right about *something* here: the violinist is indeed a person with the right to life. This suggests that it can't be generally true that if X has the right to life, causing X to die violates his rights. The reason why (4) is false in this case is that the fact that the violinist has a right to life *does not give him the right to use your body as a means for life support*. Since he doesn't have that *further* right,

disconnecting him wouldn't violate any right of his. Thomson suggests that abortions in cases where the pregnancy is due to rape should be viewed as analogous, and hence that (5) is false.

This argument says nothing about cases where pregnancy is not due to rape. Thomson does, however, offer further reasons for thinking that we should reject (5) in other cases. The first alternative case she considers is one in which the pregnancy threatens the mother's life. She develops the following variation on Violinist I to strengthen her argument by analogy:

> *Violinist II*. Like Violinist I, except that the violinist's use of your kidneys will put such a strain on you that you will probably die.

Now, even if we imagine that, prior to learning that the violinist's use of your kidneys will put such a strain on you that you will probably die, you *consented* to the operation, it seems you would be morally permitted to back out of the operation as soon as you learn this fact. So, Thomson again reasons by analogy that not only in cases of pregnancy by rape, but also in cases where pregnancy threatens the mother's life, is she permitted to abort the fetus.

Notably, nothing yet follows about whether a *third party* would be permitted to abort the fetus for the woman. For, in general, if X is permitted to do A, it doesn't follow that any arbitrary Y is permitted to do A for X. But Thomson insists that the fact that the woman's body is *hers* would enable a third party to perform the abortion for her. To support this view, she considers:

> *Stolen Coat*. Jones steals a coat from Smith to prevent himself from freezing to death. Smith will also freeze to death if he doesn't get the coat back.

It seems clear that it would be permissible for a third party to take Smith's coat back from Jones in *Stolen Coat*, even though this would lead to Jones's death. Intuitively, this is because the coat belongs to Smith. So, by analogy, since the woman's body belongs to her, a third party could abort the fetus, which is using the mother's body for life support, and which threatens her life.

What about cases where the mother's life is *not* threatened? Thomson suggests that even in these cases, abortions may be permissible *when* they would impose a substantial burden on the mother, and when it's clear that "minimal decency" would not require the mother to carry the fetus to term. In support of this conclusion, Thomson first notes that we should reject this argument:

> i. A has a right to life.
> ii. Using X is the only way to save A's life.
> ---
> iii. Therefore, A has a right to use X.

To show that this argument is invalid, Thomson appeals to the following case:

> *Henry Fonda's Cool Hand I*. The only way that you could be saved from dying from your sickness is by having Henry Fonda's cool hand touch your fevered brow. But he is thousands of miles away, and it would burden him to travel to you.

It's clearly false that you have the right to the touch of Henry Fonda's cool hand in this case. But it is still true that you have the right to life, and that being given the touch of Henry Fonda's cool hand is the only thing that could save your life. So, (i) and (ii) can be true while (iii) is false. Indeed, if we imagine that, in *Henry Fonda's Cool Hand II*, Henry Fonda is not thousands of miles

away, but just across the room, it *still* doesn't seem that you'd have a *right* to the use of his hand, though he *ought* to give it to you, and though he'd be a really crappy person not to do so. So, Thomson insists that we should also reject the following piece of reasoning:

a. A ought to give X to B, since X will save B's life.

---

b. B has the right to be given X by A, since X will save B's life.

So, by analogy, Thomson suggests that if carrying a fetus to term would impose a substantial burden on the mother, it may be permissible for her to abort it. And she also suggests that although, if carrying the fetus to term wouldn't impose any burden, it may be impermissible to abort the fetus, this is *not* due to the (putative) fact that the fetus *has the right to the use of her body*.

Now, when exactly *does* Thomson think that abortion isn't permissible? It is, I think, fair to view her as conceding that if the mother becomes pregnant *via* voluntary sex in full knowledge of a significant likelihood that it would result in pregnancy, and if carrying the fetus to term would not impose any substantial burden on the mother, it is impermissible to abort the fetus. Here the condition that it must be known to be *likely* that the sex act will result in pregnancy is important. Thomson argues for it by analogy with this exceedingly whimsical case:

> *People Seeds*. People seeds are drifting around that grow in carpets and upholstery. You install mesh screens in your windows to prevent them from drifting into your house. But, against the odds, one drifts in and takes root.

Thomson takes it that even though what resulted in the person-seed's getting into your house was a voluntary act of yours, it doesn't follow that you have any obligation to allow the seed to grow. This is because you took precautions to vastly reduce the probability that this would happen. By analogy, if a pregnancy happens in spite of the use of effective contraceptives, it doesn't automatically follow that the mother has an obligation to carry the fetus to term.

Her ultimate conclusion is that abortion is sometimes permissible and sometimes not. What distinguishes between the cases is whether it would be "minimally decent" of the mother to allow her body to be used as a means of life support. In cases of rape, threat or burden to the mother, or improbable unwanted pregnancy, minimal decency does not require this. And even when minimal decency does require an abortion not to be performed, this is not, or at least not principally, due to the fact that the fetus has a right to the use of the mother's body.

### 3.3. *Objections to Thomson*

One objection to Thomson's argument is known as the *Responsibility Objection*. According to this objection, in all cases other than cases of pregnancy by rape, the mother is responsible for the existence of the fetus. But, in many of Thomson's cases, it is not plausible that the person who kills is *responsible for the other person's need for aid*. This is very clear in both Violinist I and Violinist II. So, given the disanalogy, why think that conclusions from her cases can be extended?

Thomson's only reply to this objection turned crucially on the thought that, in cases where an effective contraceptive is used, the fact that the resulting pregnancy was unlikely frees the mother from being responsible in any *relevant sense* for the existence of the fetus or its need for aid. But people in the literature object to her appeal to improbability by noting our intuitions about the following kind of case:

> *The Cautious Hunter*. A hunter takes every reasonable precaution to avoid shooting innocent bystanders, and the objective probability of his hitting one is in fact extremely low. But the improbable does happen, and the hunter ends up shooting an innocent bystander by accident. The bystander needs a blood transfusion to survive, and the hunter has the right blood type.

Here, in spite of the improbability of the accident, the hunter does seem to have a duty to provide the transfusion, and so is responsible for the bystander's need for aid.

But this quip is a bit quick, since there is a distinction between *being responsible for someone's existence (or continued existence)*, and *being responsible for a need for aid that inevitably accompanies their existence (or continued existence)*. To see this, consider the difference between the following two cases:

> *Imperfect Drug*. A famous violinist has contracted a rare disease. The only cure for the disease is a pill that has an unfortunate side-effect: ten years after taking the pill, the violinist will likely end up with a kidney ailment which could be cured by your hooking him up to your kidneys for several months. You give the violinist the pill.
>
> *Malpractice*. Like *Imperfect Drug*, except that there are two pills, only one of which has the bad side-effect. You give the violinist the pill with the bad side-effect.

While you are responsible for the violinist's continued existence in both cases, you only have the later responsibility to hook the violinist up to your kidneys in *Malpractice*. So, if you cause someone to exist or continue to exist, and the only way to do this also makes them require aid, you do not thereby acquire the responsibility to provide that aid.

Another objection is the *Parental Bond Objection*. In none of Thomson's cases is there a biological relationship between the person in need and the person who could provide the aid. But surely there's some intuitive plausibility to the idea that the fact that the woman is the fetus's *mother* gives her a special reason to attend to its need for aid. Thomson's only reply to this objection is a flat-footed denial of the intuition, and this, to say the least, is extremely unsatisfactory.

A final objection is the *Killing vs. Letting Die Objection*. In Thomson's key cases, the candidate provider of aid only *lets the person in need of aid allow to die*. But, according to the objection, the fetus is killed in abortions. And there is a moral difference between killing and letting die. Thomson does reply to this objection with her Growing Child case, but a simpler response is that it is false that all abortions require the fetus to be killed. This is not true of hysterotomy abortions. So, at best, the objection simply shows that abortive practices should be changed, not that abortion is *per se* impermissible.

**References**

McMahan, Jeff. 2002. The Ethics of Killing. Oxford: Oxford University Press.

Parfit, Derek. Forthcoming. On What Matters. Oxford: Oxford University Press.

Rachels, James. 1986. The End of Life. Oxford: Oxford University Press.

Singer, Peter. 2005. "Ethics and Intuitions." *The Journal of Ethics* 9: 331 – 352.

Thomson, Judith. 1971. "A Defense of Abortion." *Philosophy and Public Affairs* 1.1: 47 – 66.

Thomson, Judith. 1985. "The Trolley Problem." *The Yale Law Journal* 94.6: 1395 – 1415.

Thomson, Judith. 1991. "Self-Defense." *Philosophy and Public Affairs* 20.4: 283-310.