INTRODUCTION TO PHILOSOPHY

JULIUS SCHOENHERR

Table of Contents

Dualism and causal exclusion

Concepts to know: dualism (substance and property), physicalism, causation, constitution, Leibniz's law.

Arguments to know: mind-body problem, argument for dualism, first exclusion argument (with possible solutions), second exclusion argument

There are the two metaphysical claims about the mind we will be looking at. ("Metaphysical" just means "concerned with the ultimate nature of things.")

> **Dualism**. There are two different kinds of things: mind and matter.
>
> **Physicalism**. There is only one kind of thing: physical matter. There is nothing over and above physical matter.

While these are claims about the mind quite specifically, ultimately, they bear on a much broader question about the ultimate nature of our universe:

> **The big metaphysical question.** What is the ultimate nature of our universe? Is everything physical, or is there more than just physical stuff?

Basic argument for dualism

You might wonder: why focus on the mind when trying to ask this question? The reason is that the mind is the best candidate for a non-physical thing. The reason for this is that mind and matter just seem so different, as Searle indicates, reflecting on Descartes. Here are some intuitive differences:

> **Some properties of the mind** (see Searle's text)**.** Thinking, known directly, free, indivisible, indestructible.
>
> **Some properties of matter** (see Searle's text)**.** Extension, known indirectly, determined, infinitely divisible, destructible.

We can add a few more:

> **Another alleged difference – intentionality**. Mental states are *about* things. For instance, a thought about my mother is *about my mother*; Looking outside the window, seeing a tree is a perceptual state *about the tree*. Physical states (e.g., neurons) are not about anything.
>
> **Another alleged difference – subjectivity**. Conscious experience is experienced from someone's perspective. For instance, seeing red is *my* experience. Physical facts, however,

are not subjective. Thomas Nagel famously argued for this in his paper "What is it like to be a bat?".[1]

With these seeming differences in hand, we are well on the way to craft an argument for dualism. But, first, we need a rather uncontroversial principle to state this argument:

> **Leibniz's law (aka the 'indiscernibility of identity').** If two things are identical, then they share all their properties. If things do not share all their properties, then they are not identical. ($a = b \longrightarrow \forall F(Fa \Leftrightarrow Fb)$)

This principle is fairly uncontroversial when applied to things *at a single time*. When applied to things across time it is not a plausible principle: I am the same person that I was 1 year ago, but some things about me have changed. However, *at this time*, I must surely share all my properties with myself. (Also, you don't need to know the formalized version of this; it's just for those of you who have a knack for logic.)

Now we can state our argument for dualism:

> **P1. Leibniz's law.** If two things are identical, then they share all their properties. If things do not share all their properties, then they are not identical. ($a = b \longrightarrow \forall F(Fa \Leftrightarrow Fb)$)
>
> **P2. Difference**. Mind and body do not share all their properties.
>
> **C. Dualism.** Therefore, mind and body are not identical.

Now, this argument is simple and powerful and to disprove it, we, as philosophers, have to explain away the seeming difference between our minds and our brains. Ultimately, we need to argue that brains *do think*, that either brains are free too or minds are not free either, that minds can be destroyed if brains can be destroyed, etc. This is hard work and spans many disciplines of philosophy. Here, we can only touch upon these issues.

Before moving on, let me tell you about two kinds of dualism:

> **Substance dualism.** The mind has a non-physical substance.
>
> **Property dualism.** The mind has non-physical features or, which is the same, properties.

This difference might not seem easy to understand because it's not easy to understand what exactly a substance is. The important insight, however, is simply that properties must be instantiated in something: a property is always a property of some kind of object. A substance, by contrast, can exist independently. The difference between both kinds of dualism, thus, is this: substance dualists thinks that the mind can exist independently (without the body); property dualists, by contrast, think that the mind, although non-physical in nature, needs the body to exist. The mind, as it were, is a non-physical feature *of the body*. So, just to illustrate, if you are religious and you think the mind survives the death of the body, you must be a *substance dualist*.

---

[1] Nagel, Thomas. "What is it like to be a bat?." In *The language and thought series*, pp. 159-168. Harvard University Press, 1980.

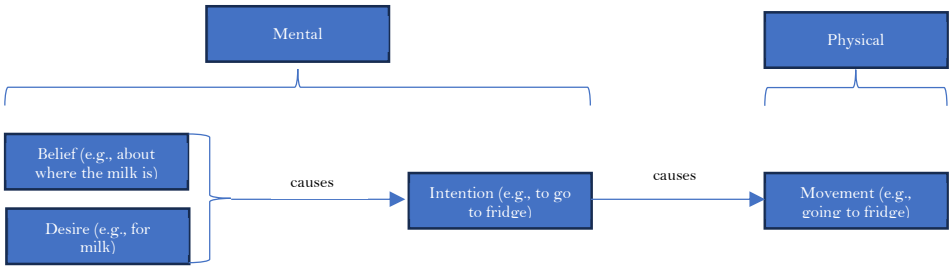
Now, you might ask: why not just accept dualism since mind and body seem so different? Here is a powerful argument against it:

> **The mind-body problem.** If dualism is true, then it is really difficult to understand how mind and body could interact. But the mind and the body do seem to interact!

Constitutive vs. causal explanations

Before we move on, to look at these arguments, it's time for some more conceptual distinctions; these are tools that will help us understand the arguments that follow:

> **Causal** mind body connections. We ordinarily think that the mind can cause the body to move. For instance, I might *want* some milk and *believe* that there is milk in the fridge. I consequently form an *intention* (i.e., a plan) that causes my body to move to the fridge. In this case, mental states such as beliefs, desires, and intentions cause the body to move.



> There are two important features of causation to keep in mind:
>
> > **Temporal structure**. Causation moves forward in time; that is, when A causes B, then A precedes (i.e., comes earlier in time than) B.
> >
> > **No self-causation.** When A causes B, then A is different from B. Nothing can cause itself.

Causal explanations – that is, explaining why something exists by looking at its causes – can be contrasted with constitutive explanations:

> **Constitutive** mind body connections. We think that the body (i.e., the brain in this case) somehow brings about our mental states. (Substance dualists will, of course, deny this.) Mental states (intentions, beliefs, desires, experiences) come into existence with the body and go out of existence when the body goes out of existence. There are a few things to know about these 'constitutive' connections.
>
> > **Language**. We say that A constitutes B (e.g., for instance, the brain constitutes the mind). But you can find other terms used; for instance, The brain realizes/ grounds/ metaphysically explains/ necessitates/ metaphysically determines the mind. They all mean the same thing.

Contemporaneous temporal structure. If A constitutes B, then A and B happen *at the same time.*

Example – the earth's axis. The earth grounds or constitutes its axis. This means that facts about the axis are the way they are *because* of facts about the earth. However, this explanation is not causal. For instance, it's not the case that, first, there is the earth, and *then*, some time later, the axis comes into existence. Rather, the axis comes into existence *at the same time* the earth comes into existence.

Brain-to-mind connection is constitutive, not causal. The mind brings about the brain in a non-causal (i.e., constitutive) way. If the brain brought about the mind causally, then we would already know that dualism is true, because cause and effect are different (see above). To illustrate the non-causal nature, imagine this: a person has a thought at a certain point in time, t1, and this person gets destroyed completely at t1. When does her thought go out of existence? At t1, or later, at t2? The correct answer is, supposedly, at t1. The thought doesn't hang around for a little bit after the brain is destroyed. If this is true, then the way the brain brings about the mind cannot be causal (because of what we said about the temporal structure of causation).

These conceptual distinctions – between causation and constitution – are very fundamental and we will encounter them throughout the course.

The first exclusion argument

We're now ready to state the first argument against dualism:

First exclusion argument

P1. Physical closure. Every physical event has a sufficient physical cause.

P2. No overdetermination. Nothing is systematically overdetermined.

P3. Distinctness. Mental events (e.g., what's going on in my mind) and physical events (e.g., what's going on in my brain) are distinct.

C. Exclusion. My (physical) bodily movements cannot be caused by my mind.

Let me quickly illustrate the individual lines of this argument:

Physical closure. This is something that we believe through scientific inquiry. Every physical event can be sufficiently causally explained by a prior physical event.

(Of course, many of you might want to ask "but what about quantum physics?" This is a valid and relevant question. Unfortunately, it's too complicated to discuss here. Further questions in this direction are: if quantum causation is not deterministic, why should mental causation be deterministic? Does non-deterministic quantum causation impact relatively coarse-grained events such as the behavior of neurons? Would we want to embrace the idea that mental causation is non-deterministic? Is there any positive,

empirical evidence for quantum causation making a difference to mental causation, or are we just proposing this because we're out of other options?) I put this in parenthesis because you don't need to know about it. It's just for those of you who are curious.

No overdetermination. We generally think that effects are not *systematically* overdetermined. This means that there are generally not two sufficient causes for an effect. Of course, *sometimes*, but just sometimes, there might two causes for the same effect. Here is a case: two shooters shoot a victim at exactly the same time. Each shot is sufficient to kill the victim. In this case, the victim's death seems to be overdetermined. Although we think that overdetermination can happen, we don't think it happens a lot, unless there is a special explanation. If dualism were true, then mind and brain would be distinct but *every action (or close to every action) would be overdetermined by a physical cause and a mental cause*. This is not plausible.

Distinctness. This is just the assumption of dualism.

How can we solve this problem?

Possible (bad) solution 1 - Epiphenomenalism. Embracing the conclusion: the mind does not cause anything outside the mind.

Objection 1 – from evolution. If the mind does not do anything, then it seems unlikely that we would evolve to have it.

Objection 2 – input output asymmetry. If we accept that the mind can be causally affected by the environment, for instance, when we perceive the environment, it seems hard to accept that the mind could not, in turn, affect the environment.

Possible (bad) solution 2 – preestablished harmony. Premise 2 is false: the body and the mind simply cause things harmony. Bodily movement *is* overdetermined.k

Objection. Although historically, you can find philosophers who have this view, it does seem unmotivated.

Possible (good) solution 3 – identity theory. Premise 3 is false: mind and body are not distinct, so mind and body can both cause without overdetermination.

This solution is indeed interesting and we'll look at it in greater detail.

The second exclusion argument

There is a second exclusion argument that employs the idea that the brain constitutes the mind.

Second exclusion argument[2]

[2] Check out Kim's (2005) book "Physicalism or something near enough" for detailed versions of these exclusion arguments.

P1. **Constitution.** Every mental event is grounded in (constituted by or realized by) a physical event.

P2. **Dualism.** Mental events are not identical to physical events.

P3. **No overdetermination.** Nothing is systematically overdetermined.

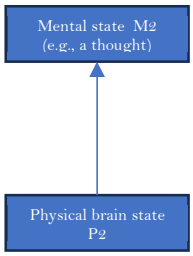
C1. Mental events E_{cause} can systematically cause other mental events E_{effect} only by causing their grounds.

P4. **Causal closure.** Every physical event has a sufficient physical cause.

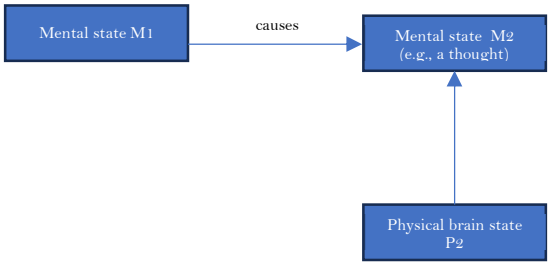
C. Mental events cannot systematically cause other mental events.

The conclusion of this argument is that thoughts cannot cause other thoughts. This is a *very* problematic conclusion, of course. Thoughts causing other thought is what we do when we think. In the argument C1 is the crux. Let me walk you through the argument:

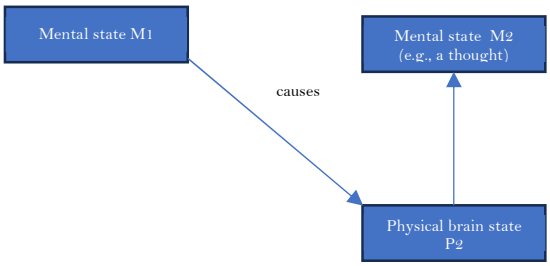
P1 says that every mental event is grounded in a physical event (e.g., a brain state). We can illustrate this as follows:



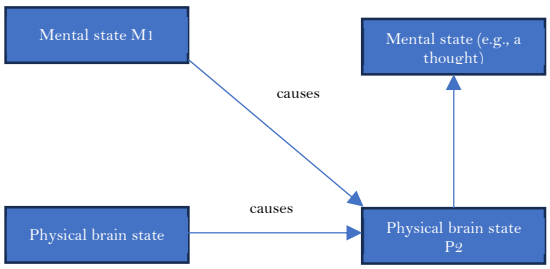
Now, suppose that there is a mental state, M1, that you think causes mental state M2:



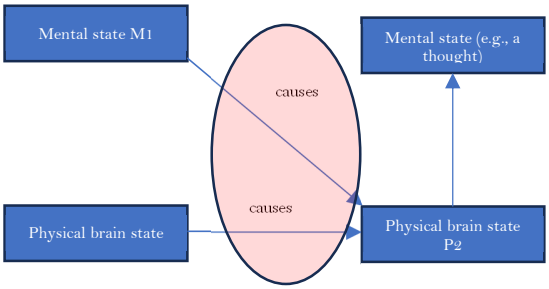
How could it cause it? Well, it seems that M1 could cause M2 only by causing P2. Here is why: The relation between P2 and M2 is roughly like the relation between the earth and its axis: the earth constitutes its axis. How could you change the axis? Well, only by making a change to the earth. You cannot change the axis independently. Because, after all, the axis is *grounded* in the earth. Therefore, it seems that this is true:



From Premise 4, we know that every physical event has a purely physical cause:



This illustrates the overdetermination problem once again. P2 has two causes:



For stuff on the exclusion argument, you can read Jaegwon Kim's book "Physicalism or something close enough." For more thoughts on dualism, I suggest you read Searle's introduction book, which is all around excellent.

Physicalism

Concepts to know: (Proper and supervenience) physicalism, supervenience, (property and event) identity theory, behaviorism, functionalism, explaining (away)

Arguments to know: argument against behaviorism, why is supervenience not enough for physicalism, machine state funtionalism

Here is, again, the statement of physicalism:

> **Physicalism**. There is only one kind of thing: physical matter. There is nothing over and above physical matter.

Before saying anything else, let's introduce one important condition of physicalism -- supervenience:

> **Supervenience.** A supervenes on B if there can be no change in A unless there is also a change in B.[3]
>
> **Supervenience physicalism**. If physicalism is true, then everything supervenes (i.e., depends) on basic physical entities.

Supervenience is a necessary condition for physicalism to be true: after all, if the mind didn't supervene on the physical stuff, then the mind could undergo changes independently of changes in the physical world. The mind would be free-floating. Now, supervenience is necessary for physicalism, but it is not sufficient. This means that it's not the whole story. Before we get to the whole story, let me clarify one further conceptual point:

You might wonder: *what are physical entities anyway?* A popular thesis for philosophers to defend is this:

> **Ideal physics.** The fundamental physical particles are those acknowledged by ideal physics, i.e., physics at the end of inquiry.[4]

You might wonder: well, maybe, separate mental entities (those that the dualist think exist) are part of our ultimate physics, at the end of inquiry. Although this might be so, you might think there is the following constraint on these ultimate entities:

> **No supervenient entities.** The entities of ultimate physics do not metaphysically depend on anything else (although they might causally depend on each other). (for the difference between causal and metaphysical dependence, please consult previous lecture notes)

So, we can at least say this much: If mental stuff (e.g., thoughts) is supervenient on physical stuff, then we know that these mental entities cannot be among the basic entities that ideal physics talks about.

Now, the statement of 'Physicalism' contains the phrase that 'There is nothing over and above physical matter.' It doesn't just say that everything just *is identical* to physical matter. Why is this slightly awkward formulation "nothing over and above" important? Well, it turns out that many physicalists think that physicalism requires merely that everything is somehow *explained* by the basic physical nature of things: whether such an explanation requires *identity*, in particular, is a further question (see below for more on identity):

---

[3] Supervenience Physicalism. O facts supervene on P facts if no two possible situations are indiscernible with respect to their P-facts while differing in their O-facts. (see Chalmers 1996, 30)

[4] See Crane, T., & Mellor, D. H. (2002). There is no Question of Physicalism. *Contemporary Materialism*, 68.

> **Proper physicalism (supervenience and explanation).** Physicalism is true if, and only if, anything that exists supervenes on *and is suitably metaphysically explainable* in terms of basic physical entities.[5]

Now, as physicalists we believe two things: first, we believe that everything supervenes on the physical nature of our world. This is just to say that there can be no change in the way things are without there also being a change in the way things are physically. Of course, given that nothing can change independent of physics, we want to know *why* this is so. As physicalists, we believe, second, that everything depends on the physical nature of the world *because of facts about physics*. That is to say, it is *no accident* that everything depends on the physical nature of our world. (Remember, when we say "metaphysical explanation" we do not mean "causal explanation.")

Identity theory

So, let's answer the explanation question: why does everything depend on the physical nature of this world, i.e., why is supervenience physicalism true? Here is one tempting answer:

> **Identity theory.** There can be no change in the world unless there is also a change in the things are physically (i.e., supervenience) *because everything is identical to some physical entity*.

To see why this seems so compelling, look at an example:

> **Superman and Clark Kent.** The fact that Clark Kent can't do anything without Superman doing it as well is explained by the fact that Superman and Clark Kent are identical. They are the same person.

So, if the identity thesis were true, then this would nicely explain why the mind supervenes on the brain. Let's apply this insight to the mental:

> **Mind-brain identity theory.** The mind and the brain are identical.

It turns out that there are at least two ways to add some precision to this formulation:

> **Mind-brain event identity theory.** Every mental event is identical to a physical event.

> **Mind-brain property identity theory.** Every mental property is identical to a physical property.

What's the difference? Let's talk about *event* identity first. Events are simply things that happen in the world (e.g., a plane taking off, Julius teaching a class, Mary celebrating her birthday). Let's consider an identity statement between a mental and a physical event:

> **Pain event identity.** John's feeling a certain pain is identical to C-fibers firing in John's brain.

[5] For more on supervenience and explanation, see, Chalmer's (1996) "The conscious mind," Chapter 2.

Now, again, events are just things that happen in the world. As such, many features can be part of the same event. To see this, consider the following event

> **'Julius giving a lecture.'** This event has many features: it involves Julius moving a certain way, saying certain things, students sitting on their chairs, a phone ringing, a spider crawling across Julius' shoe etc. These are all features that are part of this event.

The name of the event tells you about some, but certainly not all, of the features of the event. Think of events like buckets with lots of stuff in them. The name of the bucket gives you a clue about what's in the bucket, but it doesn't tell you exactly what's in it. Now consider a physical event in John's brain

> **'C-fibers firing in John's brain.'**

Suppose we think that this event is identical to one of John's mental events:

> **'C-fibers firing in John's brain' = 'John feeling a certain pain'**

Suppose it is true that these are the same events. We can now ask: what are the features that were part of this one event? We know that this event contained C-fibers firing and John's feeling pain. But we don't know that the pain involved in this event is a physical feature. It might just be that the pain was an extra, non-physical feature of this event. Remember, events are just things that happen in the world that can contain many, many different features. Our problem, thus, can be summarized like this:

> A physical event might have non-physical features (or, what is the same, properties).

Long story short: Mind-brain event identities are compatible with property dualism (i.e., the view that the mind contains purely non-physical features).

Since the problem with event identity is that these physical events might contains non-physical features, we could be tempted restate our identity theory in terms of property identity:

> **Mind-brain property identity theory.** Every mental property is identical to one, or several, physical properties.
>
> > E.g., 'heat = molecular motion' (for a non-mental example)
> >
> > E.g., 'pain = C fiber firing' (for a mental example)

Some basic facts about properties

> Properties are **general features** of the mind that can be variously instantiated.
>
> > e.g., pain is a general feature of our minds: I can have it, you can have it, etc. Think of properties as "types of things." For instance, pain in a certain type of thing.
> >
> > Identifying mental properties with physical properties means identifying general features of the mind with general features of the brain.

There is one big problem with mind-brain property identity:

> **Multiple realization.** Mental properties are multiply realizable by different physical systems.[6]

Here is an example for multiple realization: the example of thought

> **Wetware chauvinism.** Our brain produces thought, but, in principle, sophisticated computers, or aliens might also be able to think. In any case, we wouldn't want to say that aliens or computers cannot think *just because they are made of different physical stuff.*

Of course, if thought can be realized by physical structures quite unlike our brain, then thought cannot be *identical* to the physical brain itself. If two things are identical, then they cannot be separated. What thinking is cannot depend on the precise physical structure implementing thought. Maybe we should distance ourselves from the idea that mental states are identical to physical states themselves. Maybe we should, instead, embrace the idea that mental states are identical to what physical systems *do*:

> **Operational mental properties.** Mental states are not identical to physical features themselves, but, rather, mental states are particular ways that physical systems are organized.

To see what this means, imagine the following example:

> **Two printers.** The first printer is made of silicon; the second printer is made of metal. There is an obvious difference between these printers: they are made of different things. However, there is also an obvious similarity between them: they both print ink on paper when you press the 'print' button. Thus, **the physical parts are different, but the way these parts work is the same.**

Maybe mental states – e.g., thoughts, perceptions, etc. – are not identical with the physical parts themselves, but, rather, they are identical with the ways these physical parts operate or behave! Let's explore some ideas in this direction.

Behaviorism

As Searle tells us, the crudest form of behaviorism is this

> **Crude behaviorism.** The mind can be reduced to lawlike relations between sensory input and behavioral output. In short, the mind is just the behavior of the body.[7]

According to this doctrine the mind is defined in purely organizational terms: any system that produces a certain output in response to a certain input would have some relevant mental state. Here is an illustration:

---

[6] See Jerry Fodor's (1974) article "Special Sciences" for more.
[7] A classic formulation can be found in Carnap's (1932) "Psychology in Physical Language."

Believing that it is raining. The belief that it is raining = a disposition to look for an umbrella when you are confronted with rain.

As this example illustrates, behaviorism is implausible for various reasons:

No behavioral laws. There are no good candidates for pure behavioral laws, i.e., laws that do not mention mental states. "Jones will carry an umbrella if he believes that it is going to rain is only plausible if we suppose that Jones does not want to be rained on." (Searle, 53) But to "want" something is another mental state.[8]

Denying the obvious. Mental states simply seem to be real beyond input-output laws.

"The real difficulty with behaviorism, though, is that its sheer implausibility became more and more embarrassing. We do have thoughts and feelings and pains and tickles and itches, but it does not seem reasonable to suppose that these are identical with our behavior or even with our dispositions to behavior. The feeling of pain is one thing, pain behavior is something else. Behaviorism is so intuitively implausible that unsympathetic commentators often made fun of it. As early as the 1920s, I. A. Richards pointed out that to be a behaviorist you have to "feign anesthesia."9 And university lecturers have a stock repertoire of bad jokes about behaviorism. A typical joke: a behaviorist couple just after making love, he says to her "It was great for you. How was it for me?"" (Searle, 55)

Functionalism

If objections are correct, then we cannot define one mental state without appealing to other mental states. Therefore, when defining mental states in terms of behaviors of physical systems we might want to refer to inputs, outputs, *and other mental states.* This thesis is called *functionalism:*

Functionalism. Mental features such as thought are not identical to the physical structures themselves, but to the way these physical structures are organized, i.e., the way they function. In particular, a mental state S can be functionally defined in terms of its relation to sensory input, behavioral output, and other mental states.

My favorite version of functionalism is called 'machine state functionalism.' (see Ned Block, 1980 "What is Functionalism"). The basic idea is this:

Machine state functionalism. Mental states are identical to functional machine states.

Here is how it works. Remember, first, our argument for dualism that we formulated following Descartes intuitions: mind and brain cannot be identical, because they have different properties. For instance, the mind thinks, but the brain does not. Then, we said that physicalists have to argue that machine, too, can think. Now we're cashing in on this thought: according to machine state functionalism, thinking is something a machine can, in fact, do. We can describe the organization of the mind in three ways.

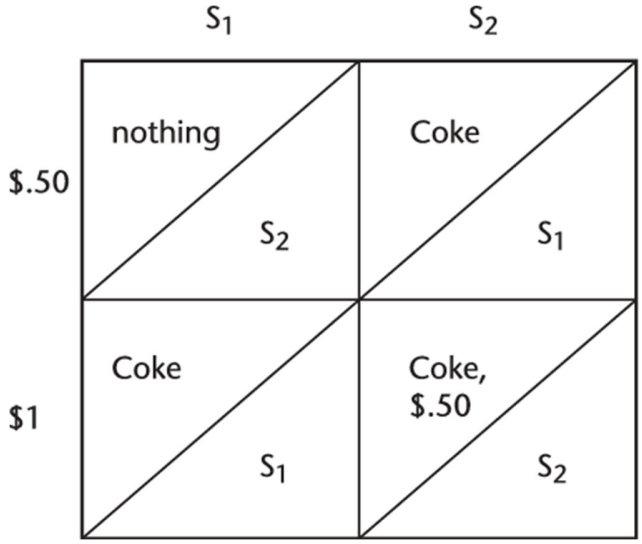
[8] The most famous version of this argument can be found in Chomsky's (1959) "A review of Skinner's Verbal Behavior."

Input. The world provides input to the mind by providing information through the senses. (e.g., mental states such as perceptual states are caused by the environment through our sensory organs). In short, through our senses, the world causes our mental states.

Internal processes. Mental states cause other mental states. That's what happens, for instance, when we think.

Output. Mental states make a causal difference to the world by causing our bodily movements.

In this sense, the mind works like a simple coke machine:



This coke machine, just like the mind, reacts to input from the environment (when people enter money), it has an internal organization (the states S1 and S2), and it produces output (it dispenses a coke when enough money has been entered. More precisely, this machine works (in part) as follows:

> Initially, the machine in S1: the state in which it dispenses a coke when $1 is entered, and it switches to S2 if someone gives it only 50 cents.
>
> If a person enters 50 cents, it switches to S2. If in S2, and a person enters 50 cents it dispenses a coke and switches back to 1.

The mind, you might think, works in similar ways. Here is an example:

> **Pain.** *"Pain [is]* the state that tends to be caused by bodily injury, to produce the belief that something is wrong with the body and the desire to be out of that state, to produce anxiety, and, in the absence of any stronger, conflicting desires, to cause wincing or moaning." (Levin 2023, SEP)

We now know how coke machines work, but is this way of talking about coke machines friendly towards our physicalist project. In particular, you might raise the following objection:

> **Non-physical machine states.** When describing the coke machine, we posited international states: S1 and S2. These states cannot be identical to the physical states of the coke machine, because any coke-machine, no matter how it is built, could have these states. So, how are these states physical.

It would be strange, of course, if we could show physicalism to be false just be describing a coke-machine. It turns out that, in describing the coke machine, we don't have to rely on any undefined, or basic, internal states. To see how this works, consider the machines states S1 and S2:

> S1 is (partly) defined as follows: the state that, if you enter 50 cents, gives you no coke and goes into state S2.
>
> S2 is (partly) defined as follows: the state that, if you enter 50 cents, gives you a coke and reverts back to S1.

Thus, S1 and S2 are interdefined: S1 is defined with reference to S2 and S2 is defined with reference to S1. Thus, there is no internal state that remains undefined and, thereby, basic.

Explaining vs. explaining away

Let's add one final, important, conceptual distinction, that between explaining and explaining away. Both behaviorism and functionalism account for mental states in terms of *doings* of physical systems, but they do so in very different ways. Functionalists explain mental states, but behaviorists explain it away.

> **Explaining vs. explaining away.** When we explain A by appeal to B, we appeal to B to show why B exists. When we explain A away by appeal to B, we appeal to B to show that B does not exist.

Behaviorists say that mental states are not real – i.e., there is no mind – but what we call the mind can be described merely by behavioral laws. Functionalists say that the mind is real – i.e. there is a mind --, but it is identical to something broadly physical: organizational properties of our brain. For this reason, it is sometimes said that behaviorists deny the mind's existence and that they try to explain its existence away; and that functionalists embrace the mind's existence and try to explain it.

Consciousness

Concepts to know: state and creature consciousness, phenomenal consciousness, blindsight

Arguments to know: zombie argument (i.e., conceivability argument), knowledge argument, Lewis' ability response

We've been thinking about physicalism: the problem of explaining how the mind can exist in a purely physical universe. Today we want to talk about consciousness, because many philosophers have argued that consciousness, quite specifically, is the aspect of the mind that is hardest to explain in purely physical terms. But let's back up a bit.

In everyday life, the word "consciousness" has many meanings.[9] For instance, we say

> "She is conscious of herself." Here conscious refers to **self-consciousness,** i.e., knowing yourself.
>
> "She lost consciousness." Here conscious refers to **being awake.**
>
> "He is conscious of his surroundings." Here conscious refers to **awareness,** i.e., being attentive.
>
> …

These are just some of the usages of the word "conscious." One basic distinction can be made between the following senses of the word "conscious":

> **Creature consciousness.** Sometimes, we use the word "conscious" to describe the creature as a whole. For instance, when we say that John lost consciousness, we mean to say that something happened to John, the entire person, namely that he lost all awareness.
>
> **State consciousness.** Sometimes, we use the word "conscious," to describe a mental state, quite specifically, not the entire person. For instance, we might say John's consciously smelled his coffee.[10]

Here, we are interested in the second sense: state consciousness. In particular, we are interested in the following sense of state consciousness:

> **Phenomenal consciousness.** A state is phenomenally conscious if there is something it is like to undergo this state.

This notion "something it is like" requires some explanation. Let's start with an intuition:

> **Piano.** Suppose someone plays a piano piece for you. This surely sounds like something: there is something it is like to hear the piano piece.

There is a famous phenomenon called "blindsight" that nicely illustrates what phenomenal consciousness is[11]:

---

[9] For more on this issue, see Ned Block's "Concepts of consciousness."

[10] You can find these distinctions in: Block, N. (2002). Some concepts of consciousness. Philosophy of mind: Classical and contemporary readings, 206-218. Or here.

[11] The classic book is Weiskrantz (1990) "Blindsight: A Case Study and Implications."

Blindsight illustrates that there is a difference between pure information processing (without consciousness), and information processing that is accompanied by consciousness: blindsighted patients process information and can, therefore, react to their environment, but what they see is not conscious: there is nothing it is like for them to see.

Note, in real life, blindsighted patients don't react to visual stimuli like non-impaired people do. In fact, they just retain very *limited* reaction to visual stimuli. However, appealing to blindsight is useful for us to illustrate the contrast between conscious and unconscious mental states: although the Blindsighted person processes visual information, their visual experiences are not phenomenally conscious.[12]

The hard problem of consciousness

We now know what a conscious mental state is. Let's talk about why these conscious states are particularly difficult to explain for physicalists.

> Chalmers on the **hard problem of consciousness**: It is undeniable that some organisms are subjects of experience. But the question of how it is that these systems are subjects of experience is perplexing. Why is it that when our cognitive systems engage in visual and auditory information-processing, we have visual or auditory experience: the quality of deep blue, the sensation of middle C? How can we explain why there is something it is like to entertain a mental image, or to experience an emotion? It is widely agreed that experience arises from a physical basis, but we have no good explanation of why and how it so arises. Why should physical processing give rise to a rich inner life at all? It seems objectively unreasonable that it should, and yet it does. If any problem qualifies as *the* problem of consciousness, it is this one. (Chalmers 1995: 212)

[12] For more philosophical applications of blindsight, see, Ned Block's (1995) "On a confusion about a function of consciousness."

So, the famous "hard problem of consciousness" is the problem of how physical processes can give rise to consciousness. Why does a bunch of neural activity give rise to experience? Why, Chalmers asks, doesn't the brain's information processing go on "in the dark"? In other words, why isn't all information processing akin to the way blindsighted people process information: pure information processing, as it were, without any experience?

Here is another way to illustrate this problem. In the previous lecture we talked about (machine state) functionalism: the view that the mind functions like a coke machine. But coke machines do not have experiences. Therefore, we ask: why: why do certain functional systems (i.e., our brains) give rise to experiences?

Now, consider again, the two pillars of physicalism:

Supervenience

Explanation

So far, we have framed the problem of consciousness as a problem *explanation:* we said that we don't know *why or how* the brain gives rise to consciousness. Many modern day dualists, excerpt for Thomas Nagel, have sought to strengthen their position, by arguing that consciousness does not even supervene on our physical brains. This means the following:

> **No supervenience claim.** It is possible that two physically identical brains differ with regard to their conscious mental states. For instance,
>
> > **Zombie claim.** It is possible that of two physically identical brains, one has conscious mental states and the other has no conscious mental states.
> >
> > **Knowledge claim.** It is possible to know all the physical facts of a brain without knowing all the consciousness facts of this brain.
> >
> > **Color inversion claim.** It is possible that of two physically identical brains, one has a red experience and the other, say, has a green experience.

The conceivability argument

Here, we will only talk about the first two of othese claims. Let's start with the claim about zombies:

> **Zombies**[13]. "[Z]ombies are beings which are physical [and functional] duplicates of us, inhabiting a world which is a physical duplicate of ours, but lacking consciousness." (Frankish, p. 2)

There is an argument that aims to show that zombies are in fact possible:

The conceivability argument

[13] For more on zombies, see, D. Chalmers' (1996) "The conscious mind," Chapter 3.

P1. Conceivability. It is conceivable that two physically identical brains differ with regard to their conscious mental states, i.e., zombies are conceivable.

P2. Link. Everything that is conceivable is possible.

C1. Possibility. It is possible that two physically identical brains differ with regard to their conscious mental states, i.e., zombies are possible.

P3. If zombies are possible then physicalism is false.

C. Physicalism is false.

Surely, P3 seems safe: if zombies really are possible, then physicalism is false. After all, if it is possible, then neither physical states nor functional states fix consciousness. Consciousness would be free-floating and, therefore, not be supervenient on the physical nature of this world. Thus, the problematic premises of this argument are P1 and P2.

Now, regarding P1, are zombies conceivable? To see what it means to be a zombie, we have to describe them in a bit more detail:

No consciousness. Although zombies only lack our conscious experiences, they have all non-conscious mental states.

Zombie thoughts. Zombies have perceptual states, thoughts, and beliefs, because these states can be functionally defined.

Therefore, in order to imagine a zombie, we have to imagine a physical-functional duplicate of ourselves – a being with our brain and all of our thoughts – that has no conscious experiences. Ned Block has called this a "*super blindsighted*" person.

Can we make sense of the idea of a zombie? There is a compelling intuition in its favor:

Lacking "this." Think of any state of which you are phenomenally conscious and ask yourself: can I make sense of the idea that a physically identical being not have had this very state? Many people find that they can make sense of this idea.

To make this more vivid. Think of a particular experience of yours, for instance, looking at a red bottle cap. Then ask yourself: is it conceivable (i.e., a coherent idea) that these neural processes in my brain could have failed to produce this conscious experience?

But there is also a compelling argument against their conceivability.

Zombie zombies. By definition, a non-conscious zombie and its conscious duplicate have all the same thoughts. A conscious person can think of a zombie, which is why a zombie must be able to think of a zombie too; a zombie zombie as it were. But what does the zombie think of when thinking of a zombie? The zombie cannot mentally *subtract*

consciousness, because, by definition, it has none. (You can find the zombie zombie argument in Peter Carruthers's (2007) paper "The phenomenal concept strategy.[14])

--- the following discussion is *not relevant* for the exam ---

Let's also consider premise 2. Is everything that is conceivable – i.e., everything that you can *think of* – also possible? This turns out to be a rather difficult question. Those of you who are not really interested in philosophy might lose motivation, I suspect, to follow what I shall now present. But since it is one of the most important topics in 20th century analytic philosophy, I will talk about it.

Many philosophers have argued that, sometimes, it takes empirical work to find out that two things are identical.[15] For instance,

Water = H2O

Heat = Molecular motion

People have always known water, but they didn't always know H2O, only empirical research, in this case in chemistry, shows that water is in fact H2O. If water is in fact H2O, then it we cannot have water without having H2O. After all, as we have been emphasizing, identical things cannot come apart. To see that it is impossible to have water without H2O, consider the following example:

> **Prank water.** Jason is a smart chemistry student. To prank his fellow classmates, he creates a substance, made of XYZ, that acts and behaves exactly like water, i.e., this substance is potable, translucent, turns solid at 0 degrees Celsius, etc.

Here, intuition has it that Jason created something that looks like water. He didn't *actually* create water. He *fooled* his classmates. Thus, unless something it is made of H2O, it is not water. In other words, it is impossible that water not be H2O.

However, to see that it is conceivable – i.e., a possibility of *thought* -- that water is not H2O, consider the following example:

> **A closer look.** Imagine that, in the year 2025, a new microscope is developed and with the help of this microscope, scientists reassess the molecular structure of various liquids (e.g., water, oil, apple juice, etc.). Indeed, these scientists make an exciting finding: water does not actually consist of H2O, but, instead, of XYZ.

As a matter of conceivability, this thought makes sense. Of course, if water is H2O, then this finding can never be made. It will forever be a possibility *of thought only*.

---

[14] Carruthers, P., & Veillet, B. (2007). The phenomenal concept strategy. *Journal of consciousness studies*, *14*(9-10), 212-236.

[15] For more on this see, for instance, Kripke, Saul A. Naming and necessity. Harvard University Press, 1980, Chapter 3.
See also Chalmers' "Does conceivability entail possibility?"

Note that 'Prank water' and 'A closer look' both talk of scenarios in which a water-like substance, XYZ, is present. However, in 'Prank water' we think of this substance as not being water, and in 'A closer look', we think of this substance as being water.

Suppose all this is true: it is conceivable that water is not H2O, but this is not possible. Then we seem to have a nice counterexample to premise 2. We might be tempted to argue:

Consciousness = certain physical features of our brain

We can conceive of consciousness without these physical features, but, just like in the water/H2O case, this might not be possible.

Let's explore the water/H2O thought a bit further: Why is it conceivable, but not possible, that water is not H2O. What is the crucial difference between conceivability and possibility? The difference lies in what is *actually the case*. In 'Prank water' we assume that water is, *in fact*, water. Given that this is true, any alternative substance that behaves superficially like water is just an imitation, not the real thing, as it were. In 'A closer look', we think of cases in which it is *not in fact true* that water is H2O; we think of a case in which it is not, and has never been true, that water is H2O. The big lesson is, thus, this: what is possible depends, in some sense, on what is actually the case. The nature of the actual world constrains what is possible. In other words, when we consider what is possible, we take the actual world for granted and ask "given that the world is actually the way it is, could it have been that …?". Conceivability, by contrast, can transcend the actual world. Here, we can even phantasize about the actual world being different from how it is. In a nutshell, what is possible seems somehow to depend on what is in fact the case; but what is conceivable does not depend on what is in fact the case; or so some philosophers have argued.

Ok, let's stop here. Philosophers have said many, many, many interesting, but also rather complicated, things about whether the case of consciousness is similar to the water/H2O case. Here, we cannot hope to decide these issues.

--- The following discussion *is* relevant to the exam, once again ---

Jackson's knowledge argument

There is a second famous philosophical argument that appeals to consciousness to show that physicalism is false (see Jackson 1986, What Mary didn't know):

> "Mary is confined to a black-and-white room, is educated through black-and-white books and through lectures relayed on black-and-white television. In this way she learns everything there is to know about the physical nature of the world. She knows all the physical facts about us and our environment, in a wide sense of 'physical' which includes everything in completed physics, chemistry, and neurophysiology, and all there is to know about the causal and relational facts consequent upon all this, including of course functional roles. If physicalism is true, she knows all there is to know. … It seems, however, that Mary does not know all there is to know. For when she is let out of the black-and-white room or given a color television, she will learn what it is like to see something red, say. This is rightly described as learning-she will not say "ho, hum."

Hence, physicalism is false. This is the knowledge argument against physicalism in one of its manifestations." (Jackson 1986, 291)

We can put this in the form of an argument:

> **The knowledge argument**
>
> **P1. Surprise.** There could be a color scientist, Mary, who knows all physical facts about color but is still surprised by consciously seeing color for the first time.
>
> **P2. Extra fact.** Mary's surprise is best explained by her learning a new fact about conscious color vision when she sees color for the first time.
>
> C. There are some non-physical facts about conscious color vision.

On the basis of this argument, Jackson believes that there are non-physical facts of experience. He calls these "qualia:"

> "I am what is sometimes known as a "**qualia freak."** I think that there are certain features of the bodily sensations especially, but also of certain perceptual experiences, which no amount of purely physical information includes. Tell me everything there is to tell about what is going on in a living brain, the kinds of states, their functional role, their relation to what goes on at other times and in other brains, and so on and so forth, and be I as clever as can be in fitting it all together, you won't have told me about the hurtfulness of pains, the itchiness of itches, pangs of jealousy, or about the characteristic experience of tasting a lemon, smelling a rose, hearing a loud noise or seeing the sky." ("Epiphenomenal qualia," p. 273)

Now, regarding **P1**: Compare the knowledge argument to the conceivability argument for just a second. In the conceivability argument we ended up thinking a lot about the step from conceivability to possibility. Although zombies are conceivable, we said, they might not be possible. But if they are possible, then physicalism is false. In this argument, we grant that the thought experiment is possible: there could be such a color scientist who knows all the physical facts but is surprised by seeing color. The possibility of this scenario is not in question.

However, in the knowledge argument it is controversial what this possibility shows. Compare, again, to the conceivability argument where it was granted that *if zombies are possible*, then materialism would be false. Whether the possibility of a scenario as described by P1 of the knowledge argument is detrimental to physicalism is *very controversial.*

Regarding P2: Many philosophers have argued that there are ways to explain Mary's surprise without granting that she learns *a new fact.* One important reply comes from Lewis' 1988 paper "What experience teaches."[16] His main thought is this:

> **Ability.** When Mary sees color for the first time, she acquires new abilities, for instance, the ability to distinguish colors by seeing them. She does not learn a new fact about color.

Lewis reasons as follows:

> **Not a matter of information.** Mary has learned all the physical facts but didn't learn what it is like to see color. However, there are no facts at all *that she could have learned*, not even non-physical facts, that would have taught her what it is like to see color. Therefore, it seems that what Mary lacks are not facts, about which she could have learned.
>
> "Black-and-white Mary may study all the parapsychology as well as all the psychophysics of color vision, but she still won't know what it's like… If there is such a thing as phenomenal information, it isn't just independent of physical information. **It's independent of every sort of information** that could be served up in lessons for the inexperienced. For it is supposed to eliminate possibilities that any amount of lessons leave open." ("What Experience Teaches," p. 289)

Now, go back to our blindsight cases. Blindsighted persons do have the ability to respond to stuff they (unconsciously) see. Therefore, we can ask: if Lewis is right, then what is the ability that Mary acquires, when seeing color for the first time, that a Blindsighted person could not (in principle) have as well?

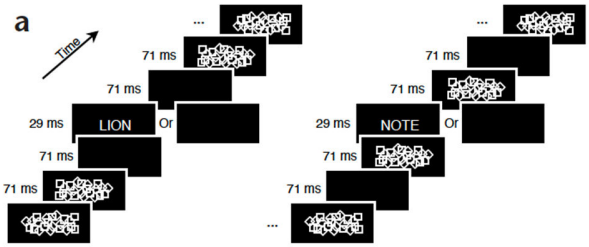
The correlates of consciousness

We can be pretty sure that certain processes in our brain give rise to consciousness. These brain processes are called the **neural correlates of consciousness,** because these are the neural structures that are correlated with consciousness. There is quite compelling evidence, mostly gathered during the past 15 years or so, that the difference between a conscious and an unconscious mental state is the conscious state's *global availability of a signal throughout the brain:*

> **Global broadcasting theory of consciousness.** A mental state is conscious if, and only if, it is globally available in the brain.

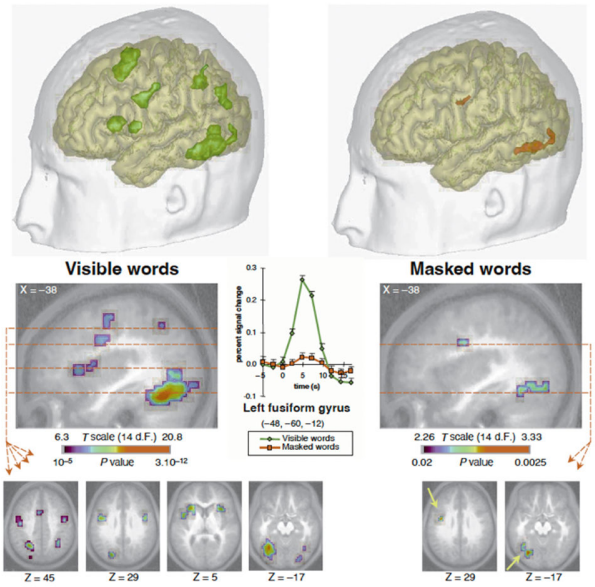
[16] Lewis, D. (1988). "What Experience Teaches", *Proceedings of the Russellian Society*, 13: 29–57; reprinted in W. G. Lycan, 1990b, 499–519, and in P. Ludlow, *et al*, 2004, 77–103

This theory goes back to Baars (1988)[17], but it has risen to popularity through much research from the past 20 years by a research group led by Stanislav Dehaene.[18] They conduct experiments such as this one: They show people pictures (i.e., *stimuli*, they are sometimes called) for a short amount of time and follow this initial picture with a second one. If the first picture is shown for a short enough time, and the second picture – which is called the *mask* – is flashy enough, then people will not realize that they actually saw the first one. In this case, the first stimulus remains what is called *subliminal*. If, by contrast, the stimulus is shown for a long enough time, then people will know they saw it. Here is an illustration of these tasks:



It turns out that *subliminally processed* stimuli (those that are not known to be seen) are processed locally: they can be correlated with activation in just a small, function-specific part of the brain. In the case of subliminal seeing, activation is registered only in the visual cortex, in the case of hearing, activation is registered only in the auditory cortex, etc. By contrast, consciously experienced mental states are registered *globally*, i.e., all over the brain. Here is an illustration of this from a paper by Dehaene et al. (2001)[19]:

[17] Baars, B. J. (1993). *A cognitive theory of consciousness*. Cambridge University Press.
[18] See, for instance, Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition*, *79*(1-2), 1-37.
[19] Dehaene, Stanislas, Lionel Naccache, Laurent Cohen, Denis Le Bihan, Jean-François Mangin, Jean-Baptiste Poline, and Denis Rivière. "Cerebral mechanisms of word masking and unconscious repetition priming." *Nature neuroscience* 4, no. 7 (2001): 752-758.



Intuitively speaking, this makes sense: when you hear something (e.g., an instrument playing) consciously, then you can remember that it played, you can act on the basis of this perception (e.g., you can go to where the music came from), you can reason about it (e.g., ask yourself whether the music is any good). Since all these abilities – memory, reasoning, action – are processed in different parts of the brain, the consciously perceived signal must be available everywhere.

As you can imagine, this kind of research is not as uncontroversial as I might have made it seem. There are philosophers (and psychologists) who do not agree that consciousness is realized by globally broadcast mental content. But investigating this further would take us too far afield.

Intentionality

Terms to know: Propositional attitude, intentionality, the causal theory of representation, Externalism about the mind

Arguments to know: Putnam's brain in the vat argument, arguments against representation as resemblance.

Propositional attitudes are a fundamental building block of our minds. Some examples of sichj attitudes are: beliefs, desires, hopes, dreams, fear, hate, speculation, intentions to name just a few. Propositional attitudes are (typically) things that take a '*that clause*'. For instance,

> He believes *that* John loves Mary
>
> I hope *that* many students will show up to the class.

This means that propositional attitudes have what is called *propositional content*. In the examples just above the propositions are "it is raining" and "students will show up."

Propositional attitude. Propositional attitudes are attitudes that embed a proposition (indicated, in English, with a *that*-clause).

(By the way, we call them propositions and not sentences because the two sentences

'He believes that John loves Mary' and

'他相信约翰爱玛丽'

Express the same propositions (or meaning), but they are different sentences. Here, in this class, we are *not* interested in language, but, rather, the contents of thoughts. Therefore, we focus on propositions to focus on the meanings these sentences express.)

Now, think back for a second to what we said about **functionalism**: that attitudes such as beliefs can be analyzed functionally. In fact, all attitudes can be defined functionally: all attitudes have their own distinct way of connecting to input, output, and other mental states. Consider the following example:

Anger

Anger is a propositional attitudes because we can say "he's angry that …". Anger is typically formed when someone injures you (input clause); anger makes you form a disposition to attack (output); and anger causes you to hope that something bad happens to the person who injured you (mental state connection). Compare, now, a different attitude:

Belief

Your belief that *p* is typically formed when you receive evidence of *p* (input); your belief causes you to be disposed to assert that *p*, say, when asked about it (output clause), and the belief that *p* usually leads to the formation of a memory that *p*. Note, however, these are just ILLUSTRATIONS, or sketches, as it were, to give you an idea of functional analysis. The actual functional analysis of 'Belief' or 'Anger' will, of course, be much more complicated.

When we talked about functionalism, we thought about how to analyze attitudes. Today, we want to think about the *contents* of these attitudes, the propositions these attitudes embed. Consider, again, the belief that John loves Mary. This belief has a particular content:

'John loves Mary.'

In other words, belief is about something: it is about John, Mary, and John's love for her. In yet other words, beliefs have intentionality, they are about things. But not just beliefs; all attitudes have content; all attitudes are about something.

Intentionality. A mental state is intentional if, and only if, it is about something/ it represents something.

Famously, the philosopher Franz Brentano remarked that *"intentionality is the mark of the mental,"* because all mental states seem to have this feature: they are about something.

> **Mark of the mental thesis.** All and only mental states have underived intentionality/aboutness.

In this statement, I have used the word "underived" because it turns out that many things have intentionality because we make it so. Here is an example:

> **Salary calculations.** You have currently 20000 yuan in your account. You want to buy something that costs 5000 yuan. Suppose you want to know how much money you have left after you spend the 5000. You get your calculator out, do the calculations and read "15000." Now, the 15000 represents the money in your account, but only *because you make it so*. By itself, the number 15000 does not represent your money.

At the beginning of our class text, Putnam gives a similar example:

> **Churchill.** An ant is crawling on a patch of sand. As it crawls, it traces a line in the sand. By pure chance the line that it traces curves and recrosses itself in such a way that it ends up looking like a recognizable caricature of Winston Churchill. Has the ant traced a picture of Winston Churchill, a picture that depicts Churchill? Most people would say, on a little reflection, that it has not. The ant, after all, has never seen Churchill, or even a picture of Churchill, and it had no intention of depicting Churchill. It simply traced a line (and even that was unintentional), a line that we can 'see as' a picture of Churchill. We can express this by saying that the line is not 'in itself ' a representation1 of anything rather than anything else.

Let's explore the idea, that mental states have intentionality, a bit further with a quick look at perceptions. This is important because it may mistakenly seem to you that perceptions – especially visual perceptions – give you some kind of direct access to the world. This is of course a mistake. Rather:

> **Mediating representations.** When you perceive (hear, see, feel) anything, say X, there is a mental state S that represents X.

We can illustrate that this is the case by appeal to hallucinations:

> **Hallucinating an oasis.** Imagine you walk through the desert when, all of the sudden, you see a nice oasis. As you go closer, you realize that there is no oasis. It was just a hallucination.

When you hallucinate, you have a mental state, a perception-like state, as of an oasis. Perceptions are similar to these hallucinations only that, in the case of perceptions, the objects that you seem to see really are there. Thus, perceptions present objects by way of mental representations.

The causal theory of representation

Now, we can ask: how does intentionality arise? Note that, when you have a thought, about anything, there is a certain brain state underlying this thought. So, we might ask, how does this brain state, a mere collection of neurons firing, come to be about something at all? Or consider words: how is it that me making certain sounds (e.g., when I say "dog") is about dogs? The most popular general answer is this:

> **Causal theories of representation.** A mental states M is about X only if there is some kind of causal relation between X and M.

Putnam believes that *non*-causal theories are *magical* and therefore implausible.

> "mental representations no more have a necessary connection with what they represent than physical representations do. The contrary supposition is a survival of magical thinking." (304)

Above, we already saw one illustration ('Churchill') of such magical thinking. Putnam gives another one. Suppose that the causal theory were false and that all that was required for intentionality was ***resemblance*** of, say, a mental imagine to what it is about:

> **Tree.** "Suppose there is a planet somewhere on which human beings have evolved (or been deposited by alien spacemen, or what have you). Suppose these humans, although otherwise like us, have never seen trees. Suppose they have never imagined trees (perhaps vegetable life exists on their planet only in the form of molds). Suppose one day a picture of a tree is accidentally dropped on their planet by a spaceship which passes on without having other contact with them. Imagine them puzzling over the picture. What in the world is this? All sorts of speculations occur to them: a building, a canopy, even an animal of some kind. But suppose they never come close to the truth.
>
> For us the picture is a representation of a tree. For these humans the picture only represents a strange object, nature and function unknown. Suppose one of them has a mental image which is exactly like one of my mental images of a tree as a result of having seen the picture. His mental image is not a representation of a tree. It is only a representation of the strange object (whatever it is) that the mysterious picture represents. Still, someone might argue that the mental image is in fact a representation of a tree, if only because the picture which caused this mental image was itself a representation of a tree to begin with. There is a causal chain from actual trees to the mental image even if it is a very strange one. But even this causal chain can be imagined absent.
>
> Suppose the 'picture of the tree' that the spaceship dropped was not really a picture of a tree, but the accidental result of some spilled paints. Even if it looked exactly like a picture of a tree, it was, in truth, no more a picture of a tree than the ant's 'caricature' of Churchill was a picture of Churchill. We can even imagine that the spaceship which dropped the 'picture' came from a planet which knew nothing of trees. Then the humans would still have mental images qualitatively identical with my image of a tree, but they would not be images which represented a tree any more than anything else." (305)

There are further reasons to reject the idea that resemblance determines what a representation is about:

> **Ambiguity.** A mental image might resemble two objects with equal accuracy but only be about one of them. Imagine, for instance that you are think of your friend Jack who happens to have a twin brother. When you think of your Jack, this thought is about *him*, even if it resembles his twin brother as well.

Symmetry. If A resembles B, then B resembles A. Therefore, if a mental image of a tree resembles trees, then trees resemble the mental image. Therefore, if aboutness were grounded in resemblance, then the image of a tree represents the tree, but trees also represent the image. This, however, is false: trees do not represent images of trees.

No images. Many think that it is an illusion that there are mental images. After all, suppose someone were to open your brain right when you see a tree. Would this person find an image of a tree in your brain? No.

Natural kind terms. Many terms for natural kinds – gold, water, elm tree, beech tree (see Putnam 309f.) refer to these kinds even if we cannot distinguish these kinds internally. For instance, for each of us, the term "gold" refers to gold, and not fool's gold; the term "beech tree" refers to beech trees and not elm trees, even if we cannot distinguish one kind from the other.

A corollary of this idea is what is called 'externalism about the mind."

Externalism about the mind. What a person's thoughts are about is not entirely determined by what is going on inside this person's body.

According to externalism, two people can have thoughts that are indistinguishable 'from the inside,' as it were, and yet these thoughts have different content, they are about different things: when you have a mental image as of a tree, then this image is about trees; when the alien has this image it is not about trees.

Now, it is a *very* general thing to say that what thoughts are about is, in some sense, determined by how these thoughts are caused. In fact, for now 50 years, philosophers have tried to make this idea more precise, but they haven't had all too much success. Here are just a few of problems that a causal theory needs to deal with:

Misrepresentation. Not every time that X causes Y, X is also about Y. Consider a case where I see a cat that looks a bit like a dog and I think "oh, there is a dog." Here, the content of my thought is DOG even though it was caused by a cat.

Inexistent objects. My thoughts about unicorns are about unicorns. However, there are no unicorns which is why unicorns could not have caused my thought.

Indirect causation. Sometimes, I just think of, say, horses when there are no horses around to cause this thought.

We could extend this list quite a bit. So, clearly, the following very simple idea is false:

Simply causal theory of representation. M is about X **if,** only if X caused M.

This view is FALSE. However, above we were more careful. We said:

Causal theories of representation. A mental states M is about X only if there is some kind of causal relation between X and M.

Here, we only say that there is "some kind of causal relation" between X and M. We didn't say that every M-state needs to be caused by X to be about X. We also didn't say that these causal relations are *sufficient* for aboutness. There are just *necessary*, i.e., causal relations have to be there, but they are not the full story. The causal theory of representation, the way we put it, is actually quite weak: it just says that without any causal relation between M and X, M cannot be about X.

Putnam's brain in a vat argument

Let's not talk about the fine details of the causal theory. This would lead us well beyond intro material and might, perhaps somewhat justifiably, bore some of you. With the help of the simple causal theory, Putnam famously argued that we can show that we are not brains in a vat. Why is this interesting?

Philosophers have, for a long time, wondered whether we can prove, know, or be certain that the external world is real. Descartes entertained this kind of thought in his famous first meditation:

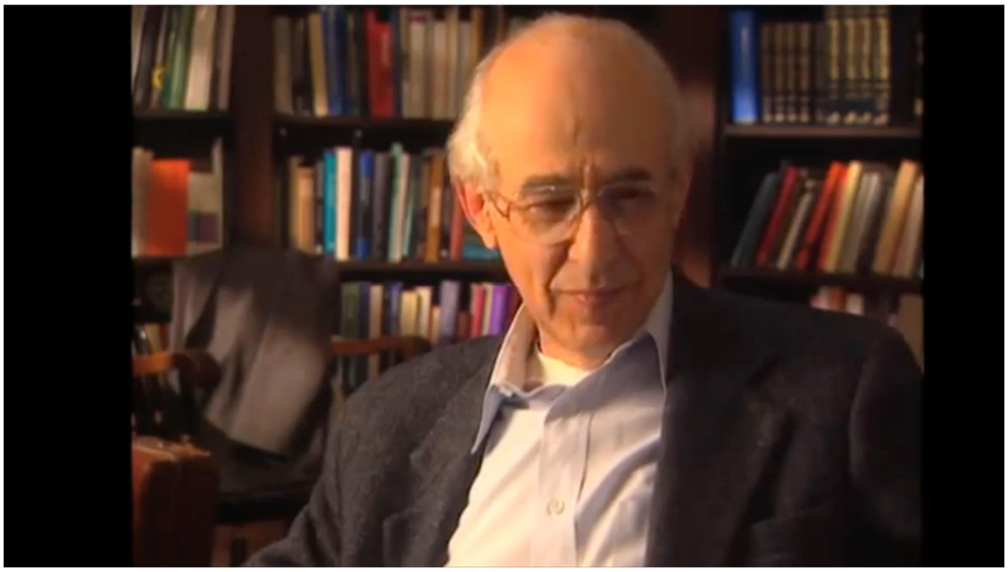
> "And yet firmly rooted in my mind is the long-standing opinion that there is an omnipotent God who made me the kind of creature that I am. How do I know that he has not brought it about that there is no earth, no sky, no extended thing, no shape, no size, no place, while at the same time ensuring that all these things appear to me to exist just as they do now? What is more, since I sometimes believe that others go astray in cases where they think they have the most perfect knowledge, may I not similarly go wrong every time I add two and three or count the sides of a square, or in some even simpler matter, if that is imaginable? But perhaps God would not have allowed me to be deceived in this way, since he is said to be supremely good. But if it were inconsistent with his goodness to have created me such that I am deceived all the time, it would seem equally foreign to his goodness to allow me to be deceived even occasionally; yet this last assertion cannot be made. (AT 7:21, CSM 2:14)
>
> I will suppose therefore that not God, who is supremely good and the source of truth, but rather some malicious demon ⌜*mauvais génie*⌝ of the utmost power and cunning has employed all his energies in order to deceive me." (quoted from Newman 2023, SEP)

The basic skeptical worry is this: if being in the real world and being deceived by an evil demon cannot be distinguished, from the inside, then it is impossible to know that I am not deceived in this way. Putnam aims to resolve the skeptical worry, proving that the external world is real.

Putnam has his own skeptical scenario:

Brains in a vat. "Imagine that a human being (you can imagine this to be yourself) has been subjected to an operation by an evil scientist. The person's brain (your brain) has been removed from the body and placed in a vat of nutrients which keeps the brain alive. The nerve endings have been connected to a super-scientific computer which causes the person whose brain it is to have the illusion that everything is perfectly normal. There seem to be people, objects, the sky, etc; but really all the person (you) is experiencing is the result of electronic impulses travelling from the computer to the nerve endings. The computer is so clever that if the person tries to raise his hand, the feedback from the computer will cause him to 'see' and 'feel' the hand being raised. Moreover, by varying the program, the evil scientist can cause the victim to 'experience' (or hallucinate) any situation or environment the evil scientist wishes. He can also obliterate the memory of the brain operation, so that the victim will seem to himself to have always been in this environment. It can even seem to the victim that he is sitting and reading these very words about the amusing but quite absurd supposition that there is an evil scientist who removes people's brains from their bodies and places them in a vat of nutrients which keep the brains alive. The nerve endings are supposed to be connected to a super-scientific computer which causes the person whose brain it is to have the illusion that …" (306)

His anti-skeptical argument can be



summarized as follows:

P1. If I am a brain in a vat, then my thought "I am a brain in a vat" is false.

P2. If I am not a brain in a vat, then my thought "I am a brain in a vat" is false.

C. My thought "I am a brain in a vat" is necessarily false.

P3. If the sentence "S" is necessarily false, then S is not the case.

C. I am not a brain in a vat.

Now, the **first premise**, P1, is pretty genius. Based on the causal theory of meaning, my VAT thoughts are not about vats, but about vats in the image, if I really were a brain in a vat. Since,

when I'm in the vat, I am not a brain in the vat-in-the-image, the thought "I am a brain in vat is false."

Premise 2 is intuitive enough.

Now, **P3** is where most people feel extreme unease: does the mere fact that the thought that "I am not a brain in a vat" is necessarily false also mean that I cannot be a brain in a vat? Think about it![20]

Now, Putnam's central philosophical tool is the causal theory of representation. It is worth emphasizing how mainstream this assumption is: the overwhelming majority of philosophers think that this theory is true: representations have meaning in virtue of the fact that these representations are causally connected to the world.

Now, I want to use Putnam's thought experiment to get us thinking about the causal theory of representation in a bit more detail. In the vat, you can't really see and there are no vats. Let's just say, for convenience, you see-in-the-image and that there are vats-in-the-image. Now, let's think about the first premise of Putnam's argument; that is, let us think about what the content of the mental state seeing-in-the-image is really about. Consider this thought experiment:

> **Brains out of the vat.** Suppose, initially, you are a brain in a vat, but one day your brain is taken out of the vat and put into a body. You now live in the actual world. Suppose that you see an apple for the very first time. While you were in the vat, all your experiences of apples were caused, not by apples, but by computer signals.

Let's just say, for convenience, that, while in the vat, you don't actually see apples, but, rather, you see-in-the-image and that there are apples-in-the-image. Compare the following two scenarios:

> **Brain in the vat situation**
>
> You see-in-the-image an apple-in-the-image. The cause of this visual-in-the-image experience, VE, is some kind of computer signal, S.
>
> **Brain out of the vat situation**
>
> You see an apple. The cause of this visual experience is a (real) apple.

Now ask yourself, is VE about S or about vats? In other words, would it be false to say "oh, while I was in the vat, everything was a great illusion, since there were no apples. Now, I realize that *this*, the real apples, was what I was thinking of all along!" Causal theories of representation, it turns out, have to deny this! Causal theorists hold that, while in the vat, your thoughts are not illusions at all: your thoughts about apples are actually about computer signals, not about apples!

---

[20] I'd like to point out one interesting feature of denying premise 3: clearly, if Putnam is right about premises 1 and 2, then it is inconceivable that I am a brain in a vat: I cannot conceive of it. However, if it is still possible, then there are some inconceivable things that are nevertheless possible. In our last class, we thought about the reverse case: whether there are some conceivable things that are not possible.

Personal identity

Concepts to know: phase vs. substance sortals, qualitative vs. quantitative identity, 4 proposals of what we are fundamentally (taken from list below), the psychological criterion

Arguments to know: teleportation cases against the psychological criterion (the brain line case), Parfit's argument from spectrum cases.

Fundamentality and persistence. Philosophical thought on what is called "personal identity" ultimately focus on this question:

> **The Fundamental Question.** What kind of thing am I, fundamentally speaking?

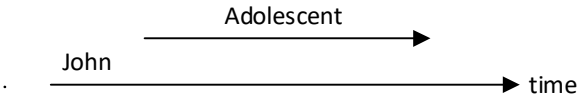
This is the question we want to understand and begin to answer. You might intuitively find the first part of this question – "What kind of thing am I?" – sensible, but be puzzled by the addition "fundamentally speaking." Here is why it is important: there are attributes that an object can lack without going out of existence. Here is an example:

> John is an adolescent, but his being an adolescent is *not* fundamental to his existence.

In fact, when John stops being an adolescent, presumably, he, John, is still there. Since John can survive losing this attribute – 'being an adolescent' –, this attribute has nothing to do with what he is, fundamentally speaking. We might say:

> **Fundamentality.** A thing, X, has some attribute, A, fundamentally only if X cannot exist without A.

Just to get you used to some terminology: attributes that are non-fundamental, that is, attributes that I only have for some time and loosing which I can survive are called *phase sortals*. We can illustrate this as follows:



We can contrast these *phase sortals* with *substance sortals*, the latter are attributes losing which I cannot survive. When I want to know what kind of thing I am, fundamentally speaking, I want to know which of my attributes are substance sortals. These substance sortals are sometimes called "identity conditions" of a person because they determine whether a particular person, say, John, exists.

Now, before we go on, it's time for a(nother) conceptual distinction:

> **Qualitative Identity.** Two objects are qualitatively the same, if, and only if, they share all their properties.
>
> **Numerical identity.** Being numerically identical means being the very same thing.

The debate on personal identity is centered around numerical identity, because we are asking what attribute makes me me? Qualitative identity can involve *different things*. Consider two indistinguishable red billiard balls. They have the same weight, color, and shape. However, they are not the same balls. There are two balls, not just one. Using our jargon, we can say that these balls are qualitatively, but not numerically, identical. Questions about personal identity about numerical identity. Questions about numerical identity are about the conditions that determine that an object is this object.

Now, when answering 'The Fundamental Question,' we can conveniently focus on *persistence*, that is, we can focus on the question of what changes I could survive and what changes, we I to undergo them, would kill me.

> **Persistence Question.** What makes me the same person over time?
>
> "Thus we might ask whether the person to whom we are speaking now is the same as the person to whom we spoke on the telephone yesterday. These are questions about identity over time. To answer such questions, we must know the criterion of personal identity: the relation between a person at one time, and a person at another time, which makes these one and the same person." (Parfit, 655)

Again, thinking about persistence – what kinds of changes I could (not) survive – helps us figure out what kind of thing I am, fundamentally speaking. For instance, since John persists (meaning, he does not die) when he ceases to be an adolescent and becomes an adult, being an adolescent is *not* fundamental to him.

Metaphysical questions vs. questions of language. Let's introduce one last thought to complete our introduction. It is natural to understand philosophical reflections about personal identity as *metaphysical reflections*, that is, as reflections that are concerned with our ultimate nature. These are reflections about *things in the world*, not reflections about *the way we use language*. Let me give you an example to illustrate what I mean;

> **Lumpl and David.** Suppose, on Monday, Michelangelo receives a lump of clay. Call this lump 'Lumpl.' Suppose that, on Tuesday, he forms a beautiful statue out of this lump which he calls 'David.' Suppose that, on Wednesday, he flattens this statue, just like a pancake. David is gone, but Lumpl, of course, is still there.

Fundamentally speaking, David is essentially a statue because he ceases to exist once he's flattened like a pancake. Lumpl is not essentially a statue. He survives being flattened like a pancake. Now suppose that you and your friend go on a trip to Florence where you look at the statue. Imagine the following conversation:

> **Friend**: "Would *it* survive being flattened like a pancake?"
>
> **You**: "Well, it depends whether you think of it as 'David' or as 'Lumpl.'
>
> **Friend**: "Well, I want to know whether *it*, this thing, independent of what I call it, would survive being flattened!"

When we talk about personal identity, we adopt the mindset of the friend: we want to know what we are, fundamentally speaking, independent of our ways to talk, or think of ourselves. Note that

it is very natural to suppose that the world in itself contains some objects (e.g., quarks, atoms, bacteria, ME): not every objection is, for their existence, dependent on how we think about this object. Thus, although Lumpl survives, and David doesn't survive, being flattened, it makes sense to suppose that there must be an answer as to whether *it* survives being fattened.[21]

You will see that, below, when we talk about Parfit, he denies exactly this: questions about who we are are largely questions about how we choose to use language. These are *not* metaphysical questions at all.

Various proposals

Now ask yourself: what kind of thing are you? When you answer this question, you should, as I suggested above, focus on persistence: what kinds of changes could you survive? Consider a very intuitive answer:

> I am a **particular human being** (fundamentally speaking)

I guess this is the most intuitive answer, the one that comes to mind most immediately. However, consider the following case:

> **Brain transplant.** Suppose your body is befallen by some illness from which you would soon die. However, you can choose to undergo a "brain transplant procedure" during which your brain would be transplanted into a new body.

Most people believe that they would survive this operation: most people believe that **brain transplantation is a way for a person to acquire a new body, not a way for a person to acquire a new brain.** Thus, it seems that I go where my brain goes. Based on these reflections, you might consider the following definition of personal identity:

> I am **a particular brain.**

Okay, to complicate things slightly, we can distinguish between *Phase sortals* – a kind of thing that you are for some time of your life – and attributes that are not part of you at all. To see the

[21] Check out Wasserman, Ryan, "Material Constitution", The Stanford Encyclopedia of Philosophy (Fall 2021 Edition), Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/fall2021/entries/material-constitution/ for much more on this.

difference, consider the idea that you are **a particular brain, fundamentally speaking**. You might worry that, if this idea is true, then there is no sense, not even in the *phase sortal* sense, in which *you are* the rest of your body at all: your body is just a kind of housing for your brain. By contrast, we ordinarily think that my body is actually *part of me*, even if it would turn out that my body is not a fundamental part of me. As you can see, the idea that I really my body is not at all a trivial assumption. If I really am just my brain, fundamentally speaking, then my body might not be part *of me* at all but just a container *for me*.[22]

Examining the history of philosophy, people have proposed many different ideas of what kind of thing we are:

A. I am **a particular human being.**
B. I am a **particular brain**.
C. I am the **thinking part of my brain.**
D. I am a series of **mental states** (e.g., memories, beliefs, desires) (the Lockean view)
E. I am a **soul**. (e.g., Cartesian Dualism)
F. I am a **human animal**. (called Animalism)
G. I am an **organism.**

H. **I do not exist.**

I grouped the last one – H – separately, because it's obviously different in one important respect: while A – G say that I do exit, only H denies that. A – G, on the other hand, are *reductive views*. These positions allege that I do exist, but that my existence consists in something else: brain, body, mental states, etc.

Reductive vs. eliminative views. In Parfit's paper, he repeatedly says that a person might not be identical to one of these things, but that the person's existence nevertheless *consists* in these attributes:

> "On this view, though a person is distinct from that person's body, and from any series of thoughts and experiences, the person's existence just consists in them. So we can call this view Constitutive Reductionism." (656)

While earlier in the semester, these statements might have seemed strange to you, you are now well-equipped to understand what Parfit means: physicalist theories of the mind and of persons do not have to embrace *identity*. There are other physicalism-friendly ways to *explain* what a mind or a person is. Although Parfit doesn't discuss these alternative ways in detail, he gives a few examples:

> "[T]he existence of a nation just consists in the existence of a group of people, on some territory, living together in certain ways. But the nation is not the same as that group of people, or that territory." (656)

[22] You can read more about this problem in Rory Madden's (2016) paper "Thinking Parts."

So, according to Parfit, people do exist, but their existence consists in something else (like A – G). He considers the last of these answers, i.e., H. Parfit considers this view, but doesn't think that it is plausible:

> "There really aren't such things as persons: there are only brains and bodies, and thoughts and other experiences.
>
> For example: Buddha has spoken thus: '0 brethren, actions do exist, and also their consequences, but the person that acts does not. ... There exists no Individual, it is only a conventional name given to a set of elements.'" (656)

The psychological criterion. Now, John Locke famously argued for option D: we are a series of mental states. In terms of persistence, we can state this view as follows:

> **Psychological connectedness.** A person remains the same person through time if and only if her psychology is connected.

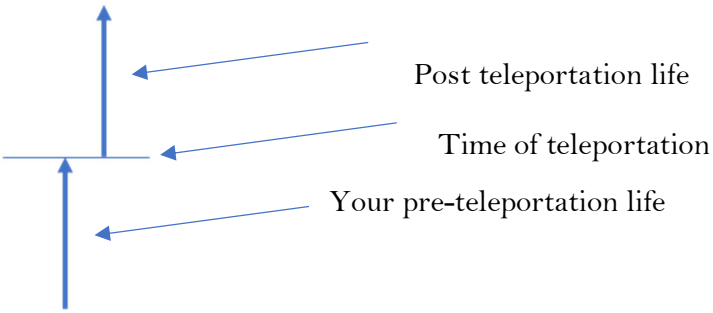
Locke believe that *memories*, and only memories, are important for my continued existence. But, on reflection, there is no real reason not to expand and say it might be *memories*, *beliefs*, *values*, *desires*, *emotions*, *etc.* that are important to who you are. To motivate this view, you can imagine, as Jeff MacMahan has, a reverse beginning of life case:

> **Reverse life**. "Imagine, for example, that in some of us the process of biological development were somehow reversed. Those to whom this happened would begin to grow younger, in biological terms. Eventually they would revert to being babies and thereafter would have to be placed in artificial wombs in order to survive. As their brains reverted to the infantile and fetal stages of their development, their mental lives would become increasingly rudimentary and would eventually disappear altogether when their brains ceased to be capable of supporting consciousness." (McMahan 2002, 29)

Ask yourself: when, in this process, do you stop existing? At least some people are tempted to say that they stop existing when enough of their mental states are gone. Of, course, this is just an intuition. You don't have to believe this. Now, there is an interesting argument for and against the psychological criterion for personal identity (see option D, above) based on the idea of teleportation.

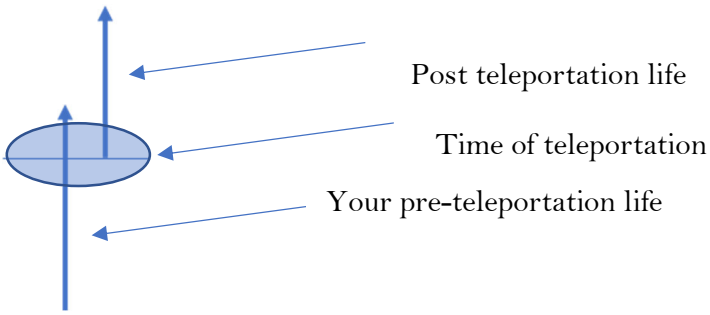
> **Teletransporter.** "We can start with some science fiction. Here on Earth, I enter the Teletransporter. When I press some button, a machine destroys my body, while recording the exact states of all my cells. The information is sent by radio to Mars, where another machine makes, out of organic materials, a perfect copy of my body. The person who wakes up on Mars seems to remember living my life up to the moment when I pressed the button, and he is in every other way just like me. Of those who have thought about such cases, some believe that it would be I who would wake up on Mars. They regard Teletransportation as merely the fastest way of travelling. Others believe that, if I chose to be Teletransported, I would be making a terrible mistake." (655)

We can illustrate this situation as follows:



Some people do believe that teletransportation does preserve personal identity. If this is true, then physical continuity, of course, does not matter for personal identity. Consider, now, an alternative to this case, in which my original body fails to be destroyed:

> **Branch line case.** "Suppose next that we believe that, even in Teletransportation, my Replica would be me. We should then consider a different version of that case, in which the Scanner would get its information without destroying my body, and my Replica would be made while I was still alive. In this version of the case, we may agree that my Replica would not be me. This may shake our view that, in the original version of case, he would be me." (658)



Why, you might ask, does *this* case tell us something about the original Teletransporter case in which the original person is destroyed? Well, if the teleported person is me in the original case, but not in the Branch Line Case, then the question whether a particular person (i.e., the teleported person) is me depends on seemingly unrelated facts, e.g., on whether some person somewhere else was destroyed. This seems implausible.

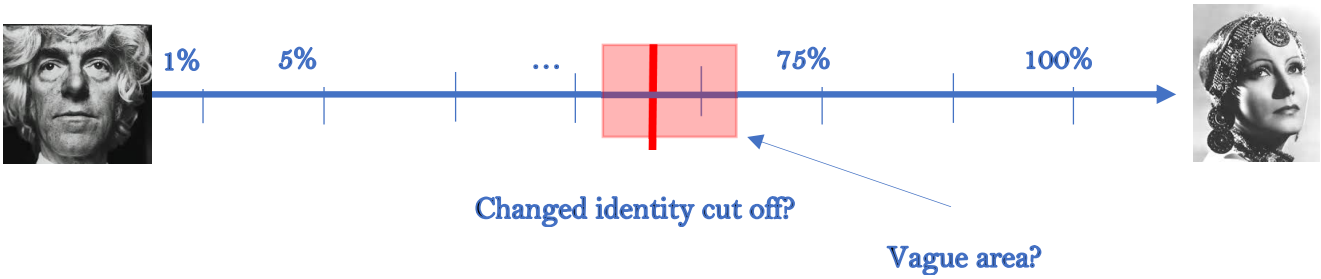
Spectrum cases and the reductionist view

Now, we considered before, albeit briefly, the idea that your continued existence consists in the continuation of enough of your body, brain, or mental states. An obvious question is: how much continuation is enough for my survival? Parfit now imagines a spectrum of increasingly thorough replacement of one's brain.

> **Combined Spectrum.** "In this second range of cases, there would be all the different degrees of both physical and psychological connectedness. The new cells would not be exactly similar. The greater the proportion of my body that would be replaced, the less like me would the resulting person be. In the case at the far end of this range, my whole

body would be destroyed, and they would make a Replica of some quite different person, such as Greta Garbo. Garbo's Replica would clearly not be me. **In the case at the near end, with no replacement, the resulting person would be me. On any view, there must be cases in between where we could not answer our question.** For simplicity, I shall consider only the Physical Spectrum, and I shall assume that, in some of the cases in this range, we cannot answer the question whether the resulting person would be me. My remarks could be transferred, with some adjustment, to the Combined Spectrum."

For illustration:



Parfit thinks that, when we reflect on such cases, we find a conflict with the following deep conviction that we have about ourselves:

> **Determinacy.** "Our identity must be determinate. We assume that, in every imaginable case, questions about our identity must have answers, which must be either, and quite simply, Yes or No." (656)

To be clear, Parfit thinks that it is *not* always determinate whether we exist. Importantly, he thinks that the existence of such vague cases – cases in which it is neither true nor false whether a person exists – show something important about personal identity: that there are no interesting facts about personal identity to begin with:

> "We can always ask, 'Would that future person be me?' But, in some of these cases,
>
> (7) This question would have no answer. It would be neither true nor false that this person would be me. And
>
> (8) This question would be empty. Even without an answer, we could know the full truth about what happened." (659)

Sorites paradox. Parfit thinks that there are simply facts about brains, mental states, bodies, etc, and that any further question of 'where I am' in this mix is purely linguistic. To illustrate this further, imagine a heap of sand from which you take a single grain of sand away. Ask yourself: "Is it still a heap?" Presumably, the answer is "yes". Then take another grain away, and another one and then another one. Does taking away a single kernel of sand ever transform a heap into a non-heap? Presumable not. However, when all the kernels are gone, there is no heap left, so the transformation must have occurred somewhere. But we've already determined that no single kernel made the difference between being a heap and not being a heap. Thus, the transformation occurred, but it seemed to have occurred nowhere. This is known as the *sorites paradox*.

To solve this problem, many accept that 'being a heap' is **vague**. In some cases, it is simply indeterminate whether something is, or is not a heap. More precisely, we might say that, in some cases, it *neither true nor false* whether an object is or is not a heap. This, in turn, might tell us something about heaps: that 'being a heap' is no further fact beyond the number of grains of sand that are piled up. Talk of "heaps" is just a way for us to talk about piled up grains of sand, but this is a matter of language, not metaphysics.

> **No Metaphysical Vagueness.** Vagueness in most cases seems to be a semantic phenomenon. 'Being a heap' is vague, because the *application of the concept 'heap' is sometimes indeterminate*. But although the application of words might be vague, *the world* is never vague.

Parfit's conclusion, thus, is that what I am, fundamentally speaking, is not an interesting question at all: it's a question about how we decide to use language. It's not a question about what kinds of things exist in the world.

Now, you might think that, if there are no deep facts about my existence – that is, if whether I exist or not is largely a matter of how we use language – we shouldn't really care about whether we live or die. Parfit surprisingly rejects this: he thinks questions about my existence or one thing, questions about what should matter to me are a different thing.

Parfit's on what matters

In his famous book *Reasons and Persons*, Parfit argues that what should matter to me is the proper connection between my mental states. Since this connection has nothing to do with personal identity, he simply calls it "Relation R":

> **Relation R.** Personal identity does not consist in psychological connectedness. Nevertheless, he defines a new term --- Relation R (215) --- that does consist in psychological connectedness.

Psychological connection does not preserve personal identity, but it preserves **what matters**. Parfit claims that Relation R is what matters. But to see how strange this view is, consider the following case:

> "One example may be the **Branch-Line Case,** where my life briefly overlaps with that of my Replica. Suppose that we believe that I and my Replica are two different people. I am

about to die, but my Replica will live for another forty years. If personal identity is what matters, I should regard my prospect here as being nearly as bad as ordinary death. But if what matters is Relation R, with any cause, I should regard this way of dying as being about as good as ordinary survival." (215)

Now, take a minute to appreciate how strange this is. Imagine yourself in this situation. Suppose you survive the teleportation, but the teleporter's radioactive radiation will cause you to die within a week. Would you be content that there is someone on Mars who's quite like you?

Knowledge

Concepts to know. TBJ definition of knowledges

Arguments to know. The Gettier problem, fake barn cases

When thinking philosophically about knowledge, we should distinguish (at least) two questions:

What is knowledge?

Do we have knowledge?

Today, we're going to think about the first of these questions, leaving the second one for another time. Notice that even if we find an answer to the first question, even, that is, if we find a suitable definition of knowledge, it might still be true that nothing satisfies this definition. In this case, we would know what knowledge is, but we would have no knowledge. During our class on intentionality, we talked about one skeptical argument: the idea that we are brains in a vat. If we are brains in a vat, then most of our beliefs about the world might in fact be false and most of the things that we *think* we know, we don't in fact know. So, philosophical skepticism is concerned with the second question.

Knowledge and truth

Now, over the years, I've heard many students say things like this:

"Knowledge is relative! 500 years ago, everybody knew that the earth was flat and now we know that it is round."

The idea behind these kinds of statements is that knowledge is some kind of *widely shared belief*. Hence, what we know changes according to what is widely believed. This view is simply false! To see why, notice, first, that knowledge entails truth:

Knowledge truth link. Knowledge entails truth.

The following statements indicate this quite well:

#"Julius didn't rob the bank, but the students know that he did"

Rather: "Julius didn't rob the bank, but the students *think* that they know he did it."

When a person is super certain about something that turns out to be false, our intuitive verdict is, not that they know, but that they *think* that they know. You might wonder: well, maybe truth is relative, too. This view is false, too. For now, you can think of truth as follows:

> **Truth as correspondence.** Truth consists in the correspondence of a proposition with the facts.

We will talk about this idea more, next time. For now, it's enough to note that facts are simply states of the world. They don't change and they do not depend on anyone's opinions or beliefs. For instance, if it is a fact that Brutus killed Cesar, then this will be forever the case and even if Brutus were to convince everyone around him that Cesar killed himself, it would still be a fact that Brutus did it. The reason why I'm stressing all this, is to show you that knowledge has nothing to do with what is commonly held to be true, or common wisdom or anything like that. Knowledge is firmly anchored in facts, as it were: Knowledge requires truth, which consists in the correspondence with the facts. Again, next time, we shall talk more about truth.

Now, earlier in our class, we talked about propositional attitudes such as hoping, wondering, believing, accepting, etc. All these, we said, are *propositional attitudes*, because they take a proposition as content. Let's, now, further distinguish between *factive* and *non-factive* attitudes. *Factive* attitudes are those that require truth. Knowledge is one of these factive attitudes:

> **Factive attitudes.** An attitude is factive if it requires truth. Examples are *seeing*, *hearing*, *knowing*.

For instance, statements such as the following sound strange, because they ascribe a factive (i.e., truth entailing) attitude but then go on to deny their truth:

> #Julius saw the student entering the classroom, but no student entered the class room.
>
> #Julius knows that his computer is out of battery, but it is not out of battery.

Knowledge and belief

Knowledge is an attitude that requires truth. But what kind of attitude is it? Most people think that it is a *belief:*

> **Knowledge belief link.** Knowledge entails belief.

The following statements indicate this quite well:

> #"Julius knows that Jane handed in the essay, but he doesn't believe that she did.
>
> Rather: Julius believes that Jane handed in the essay, in fact, he knows it.

Now, while most people do believe that knowledge entails belief, this view has been challenged by examples such as this one:

> **Unconfident examinee:** Kate is taking a history test. She had studied carefully and has been doing well on all the questions so far. She has now reached the final question, which

> reads "What year did Queen Elizabeth die?" As Kate reads this question she feels relief, since she had expected this question and memorized the answer. But before Kate can pause to recall the date, the teacher interrupts and announces that there is only one minute left. Now Kate panics. Her grip tightens around her pen. Her mind goes blank, and nothing comes to her. She feels that she can only guess. So, feeling shaken and dejected, she writes "1603"—which is of course exactly the right answer." (See Rose and Schaffer 2013, Woozley 1952, p. 155; Radford 1966)[23]

Here, it seems that the student knows the answer, but doesn't believe it. If your intuitions chime with this assessment, you might want to rephrase the belief-requirement of knowledge to something slightly weaker:

> **Knowledge dispositional belief link.** Knowledge entails the disposition to belief.[24]

Although the details of this idea are difficult, the basic idea is simple: knowledge entails that you would believe the relevant proposition if these exceptional circumstances, such as the teacher pressuring the student, are removed.

Knowledge and justification

The last basic ingredient in knowledge is *justification*. Not every true belief is knowledge. To see this, consider a person who is gullible and believes everything they are told. Of course, some of their beliefs will be true, more or less by chance, but you wouldn't say that they know these things. Knowledge is an epistemically *good* state. If a person knows something then their belief about this matter is justified:

> **Knowledge justification link.** Knowledge entails justification.

Consider a relatively mundane case:

> Julius knows that no one received an A in the final exam.

Here, it is natural to assume that there is *a reason* why I know this. If I know that no one received an A, then it seems like I don't *just* believe this. My belief must be supported by some kind of reason. For instance, I might know that no one received an A because I was the one grading the exam.

The Gettier problem

We can put these ingredients of knowledge together and arrive at the following view of knowledge:

> **TBJ analysis of knowledge.** Knowledge is justified, true belief.

---

[23] Woozley, A. D. (1952). Knowing and not knowing. Proceedings of the Aristotelian Society, 53, 151–72.
Radford, C. (1966). Knowledge—By examples. Analysis, 27, 1–11.
[24] See, Rose, D., & Schaffer, J. (2013). Knowledge entails dispositional belief. Philosophical Studies, 166, 19-50.

Now, somewhat ingeniously, Edmund Gettier found that this analysis is flawed in that it leaves something out. Gettier showed that there are certain cases of true justified belief that are *not* knowledge. Here is his first case:

Objection case 1. Consider one of Smith's beliefs:

> **Belief**. "The man who will get the job has 10 coins in his pocket"

Suppose this belief is justified because

> **Justification.** Smith has strong reason to believe that Jones will get the job, and he counted the coins in Jones' pocket and found that there is 10.

But suppose further

> **Truth.** Smith will get the job, and he happens to have 10 coins in his pocket (which Smith did not check).

Intuitively, Smith does not know that the man who will get the job has 10 coins in his pocket.

Objection case 2. Consider one of Smith's beliefs:

> **Belief**. "Jones owns a Ford, or Brown is in Boston"

Suppose this belief is justified because

> **Justification.** Smith has seen Jones with a Ford, Jones offered Smith a ride with a Ford, etc.
>
> Note: justification need not be infallible.

But suppose further

> **Truth.** Jones does not own a Ford, but Brown is in Boston.

Smith does not know that Jones owns are Ford or Brown is in Boston.

Now, since there are cases that satisfy the TBJ analysis of knowledge but that are, at least intuitively, not cases of knowledge, this definition is *too allowing*, it includes too much. In other words, it is *not sufficient* for knowledge. Thus, we might want to add an ingredient in order to make our definition of knowledge sufficient:

> **Knowledge*** Knowledge is true justified belief **plus an extra ingredient.**

Possible solutions to Gettier cases

Now, let's think about why these Gettier cases might fail to be knowledge. Consider, again, Gettier's second case:

> **Belief**. "Jones owns a Ford, or Brown is in Boston"

Justification. Smith has seen Jones with a Ford, Jones offered Smith a ride with a Ford, etc.

Truth. Jones does not own a Ford, but Brown is in Boston.

Although Smith has justification for his belief, intuitively, his justification, you might think, is **misleading**, it does not point Smith towards the truth. Smith's justification seems to point him, first and foremost, to believe that Jones owns a Ford. Once Smith believes that Jones owns a Ford, he can infer that "Jones owns a Ford, or Brown is in Boston." Somewhat more rigorously, we might reconstruct Smith's reasoning as follows:

Justification. Smith has seen Jones with a Ford, Jones offered Smith a ride with a Ford, etc.

First belief. Based on his justification, Smith comes to believe that **Jones owns a Ford.**

Now Smith can perform the following inference:

P1. Jones owns a Ford.

P2. A disjunction (or-statement) is true if at least one disjunct is true.

C. "Jones owns a Ford, or Brown is in Boston"

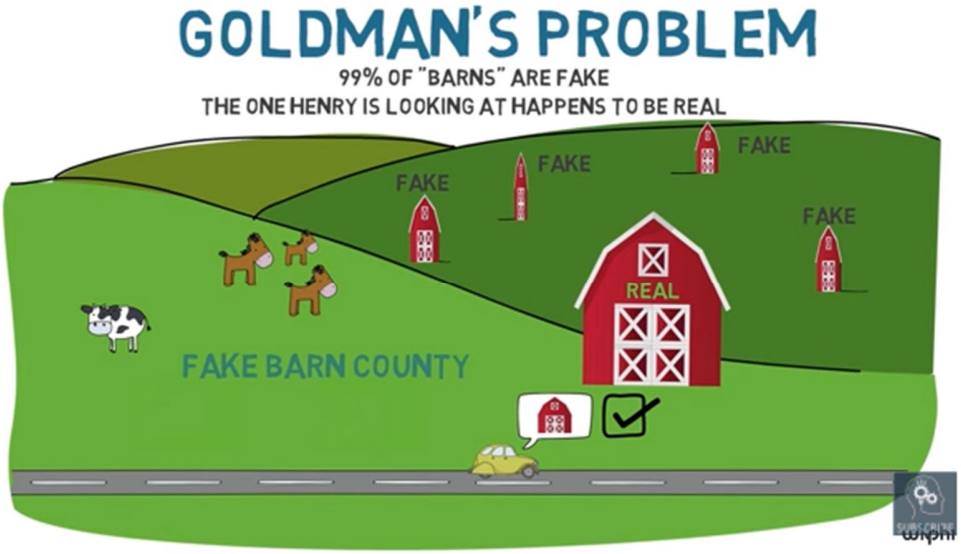
Notice, however, that P1 is false: Jones does not own a Ford. Our first idea to fix the Gettier problem might be to rule out inferences from false premises:

No false premises. If a person knows that *p*, then this person did not infer *p* from a false premise.

But Goldman (1976) construed an ingenious counterexample to this thesis, an example in which a person fails to know some true, and justified belief but where no inference from false premises seems to have occurred:

The fake barns (Goldman 1976, "Discrimination and Perceptual Knowledge", 772f.). "Consider the following example. Henry is driving in the countryside with his son. For the boy's edification Henry identifies various objects on the landscape as they come into view. "That's a cow," says Henry, "That's a tractor," "That's a silo," "That's a barn," etc. Henry has no doubt about the identity of these objects; in particular, he has no doubt that the last-mentioned object is a barn, which indeed it is. Each of the identified objects has features characteristic of its type. Moreover, each object is fully in view, Henry has excellent eyesight, and he has enough time to look at them reasonably carefully, since there is little traffic to distract him. Given this information, would we say that Henry knows that the object is a barn? Most of us would have little hesitation in saying this, so long as we were not in a certain philosophical frame of mind. Contrast our inclination here with the inclination we would have if we were given some additional information. Suppose we are told that, unknown to Henry, the district he has just entered is full of papier-mache facsimiles of barns. These facsimiles look from the road exactly like barns, but are really just facades, without back walls or interiors, quite incapable of being used

as barns. They are so cleverly constructed that travelers invariably mistake them for barns. Having just entered the district, Henry has not encountered any facsimiles; the object he sees is a genuine barn. But if the object on that site were a facsimile, Henry would mistake it for a barn. Given this new information, we would be strongly inclined to withdraw the claim that Henry knows the object is a barn.



Ok, funny videos aside, the important point about this case is that in Goldman's example, the relevant belief – "there is a barn" – is not inferred from a falsehood. To see this, a natural way to reconstruct Henry's reasoning is as follows:

> **P1.** This thing in front of me looks like a barn. (Henry looks at a real barn.)
>
> **P2.** If something looks like a barn, then it is likely a barn.
>
> **C.** This thing in front of me is a barn.

Reflecting on fake barn cases, we might be tempted to try a different solution to solve the Gettier problem. Maybe the problem with Gettier cases is not that it involves reasoning from false premises, but, rather, that these cases involve luck. To illustrate, consider, again, the first Gettier case.

> **Belief**. "The man who will get the job has 10 coins in his pocket"

As it happens, Smith has 10 coins in his pocket. But Smith got lucky! He didn't count how many coins he himself had in his pocket. So maybe we should add a 'no luck' clause to our definition of knowledge:

> **No luck.** Knowledge requires that the relevant proposition is not true as a matter of luck.

It turns out that luck is a really complicated matter, and, as it also turns out, there are many cases where luck does not undermine knowledge at all:

> **Burglar**. Smith believes that his keys are on the table. He put them there an hour ago, right before he left for work. Without his knowledge, a burglar who intended to steal his keys was about to break into his house. However, lucky to Smith, the Burglar got hit by a lightning bolt right before he got to Smith's house. Smith got lucky. His keys are still there.[25]
>
> **Robbery**. During a bank heist, the robber's mask momentarily falls off, revealing to the shocked teller that the criminal is none other than the bank president. In this scenario, while the teller is lucky to have seen this revealing moment, she undoubtedly knows that the bank president is the culprit. (Nozick 1981[26]

These cases show that the relationship between luck and knowledge is complicated: in some cases, luck undermines knowledge, but in other cases it does not. Let's leave it at that.

The value of knowledge

You might legitimately wonder why philosophers focus on knowledge so much. Why does knowledge have such a special place when analyzing how we think about the world. What role does knowledge play in our lives that, say, true, justified belief cannot play? This is indeed a tough question to which no one really knows the answer. Although it is hard to say why

[25] Cases of this kind are discussed in Comesana (2005) "Unsafe Knowledge."
[26] Nozick R. Knowledge and skepticism. In: Nozick R, editor. Philosophical explanations. Harvard University Press; 1981. pp. 167–288

knowledge is better than, say, tbj, we can point to some of the roles that knowledge seems to play in our lives:

Knowledge and assertion. First, it is popular to hold that knowledge plays some kind of special role when evaluating the correctness of assertion. To see this, consider the following case from Hawthorne (2004):

Hawthorne's *Knowledge and lotteries* (2004, chapter 1):

> **Lottery.** You buy a lottery ticket. The chances of winning are really small (let's say 1/10 000 000). Hence, you are justified in believing that your ticket will not win. Without even checking the ticket, may you assert to someone "Look, I have a loosing ticket!"

Intuitively, the answer is "no!" But why not? It's overwhelmingly likely that your ticket loses. How would you criticize someone who does make this assertion? Intuitively, I would say "how can you just say that? You don't *know* that this ticket loses." Knowledge seems to be the standard for assertion which is why we criticize people who assert without knowing.

> **Knowledge norm of assertion.** Knowledge is the norm of assertion.

Knowledge and inquiry. A second area of our lives where knowledge seems to play some role is *inquiry*, the type of thing we are doing when we are trying to figure something out. Consider an example:

> **Exam date**. Suppose you are trying to figure out the date of the final exam for this class. Suppose you can't find it on WeChat and your friends don't know either. So, you text your TAs, etc. When should you stop inquiring?

It's intuitive to say that you should stop inquiring when you *know the answer*, that is, if you keep inquiring for the exam date even though you clearly do know the answer (maybe your TA told you and you found the date online), you start seeming more and more irrational. Inquiry beyond knowledge seems strange and out of place.

> **Knowledge aim of inquiry.** Inquiry aims at knowledge/

The idea is that inquiring whether something, *p*, *is true*, means *wanting to know* whether *p* is true, and that you should stop inquiring once you do know that *p*. This is a claim that has been defended most powerfully by the philosopher Jane Friedman.

I should stress, finall, that these ideas – especially the ideas concerning the role knowledge plays in inquiry and assertion – are *heavily* contested. However, I presented these views to you to give you an idea of how one might argue for the importance of knowledge.

Truth and facts

Concepts to know: The combined view, The problem with negative facts,

In Russell's short piece "Do Facts Make True Whatever Is True?", he gives a remarkable overview over the basic metaphysical structure of our world with "facts" at the center. In this

course, we have been focusing, in some way or other, on the basic structure of the world: what is the world like fundamentally speaking. For a while, we looked just at concrete things and asked whether all these concrete things are physical. Now, in this last class, we're changing our perspective a little bit and ask: are all things in this world simple or complex, and are they all concrete or are there also non-concrete objects that we might call "facts"?

> "Facts are … something you have to take account of if you are going to give a complete account of the world. You cannot do that by merely enumerating the particular things that are in it: you must also mention the relation of these things, and their properties, and so forth, all of which are facts, so that facts certainly belong to an account of the objective world …" (105)

Russell begins by telling us that "the world contains facts" (103). Furthermore, he emphasizes that there are all kinds of facts such as:

> **Mathematical facts.** '2+2=4'
>
> **Particular facts**. 'There is a bottle in my hand.'
>
> **General facts.** 'All humans are mortal.'
>
> **Positive facts.** 'Socrates was alive.'
>
> **Negative facts.** 'Socrates is not alive.'

The first, and most important, thing to notice about facts is that they are part of our world. Russell says

> "The world contains facts" (103).

This idea is naturally contrasted with the idea that the world is just a collection of **objects**:

> ☺ **World of facts.** Fundamentally speaking, the world contains facts.
>
> ☹ **World of objects.** Fundamentally speaking, the world is just a collection of things.

Russell believes the first of these views.

> "The outer world … is not completely described by a lot of particulars, but that you must also take account of these things that I call facts."

But what are facts? According to Russell, it seems, facts are complex in that they contain objects and their properties:

> "We express a fact, for example, when we say that a certain thing has a certain property, or that it has a certain relation to another thing; but the thing which has the property or the relation is not what I call a 'fact'." (104)

Now, Russell seems to think that facts are complex in that they consist in things having certain properties, or standing in certain relations. For instance:

Fact: 'Julius taught the intro class on April 16, 2024.'

This fact is complex in that it says of Julius that he has a certain property: that of giving a lecture.

Facts vs. events. We should note, maybe in slight opposition to Russell, that facts should be distinguished from concrete things having certain properties. To see this, we can contrast facts from *events* both of which are complex and describe things and their properties. Consider the corresponding event:

Event: 'Julius' teaching the intro class on April 16, 2024.'

Now ask yourself: when did this event end? The answer is that it ended at 5 PM, on April 16, 2024. And now ask yourself: when did the (corresponding) fact end? Here you will probably say: "there is no such end. It will always be a fact that Julius taught the class on April 16, 2024." You would be right in saying this.

Facts are part of the world, but, in contrast to particulars, they do not exist in time and space. They have no end, and sometimes no beginning. They keep existing even if the particulars go out of existence. For this reason, philosophers sometimes say that facts are **abstract objects.**

Now, although facts are not themselves particular things, they are, in some sense, *grounded* in particular things. Consider the following fact:

'All humans are mortal.'

Now ask yourself: why is this a fact? The answer, presumable, is that this is a fact because all men are in fact mortal. That is, because a certain kind of thing – that we call 'humans' – have a certain features – that of being mortal. Thus, concrete (e.g., humans) and their properties (e.g., being mortal) *ground* these facts, but these facts are not identical to these particular things: facts are abstract objects outside of space and time.

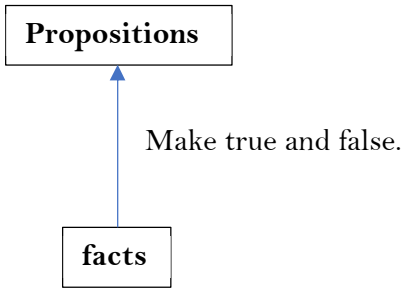
Function of facts. The main function of facts is to make propositions true. Consider such a proposition:

True Proposition. ⌜Julius taught the intro class on April 16, 2024.⌝

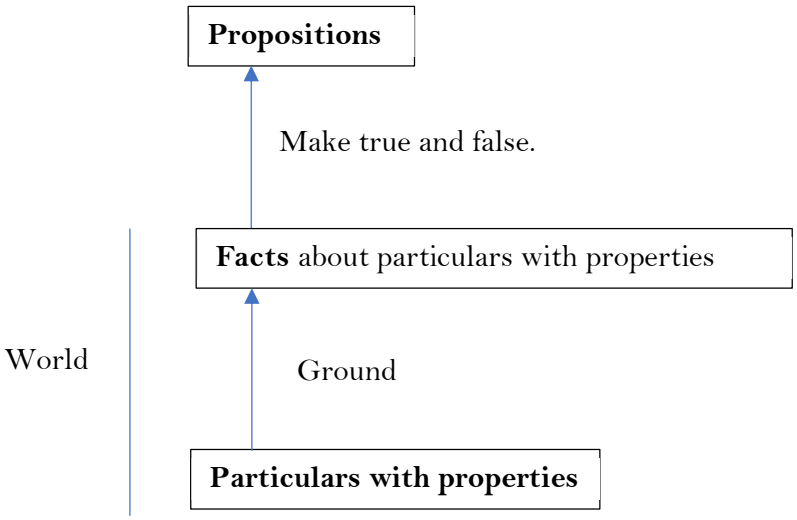
I write these square brackets "⌜⌝" just to indicate that now we're talking about propositions, not facts. This proposition is true because what it says is a fact. We can furthermore, maybe a bit more controversially, say that facts can also make propositions false. Consider the following proposition:

False Proposition. ⌜Julius taught the intro class on April **20**, 2024.⌝

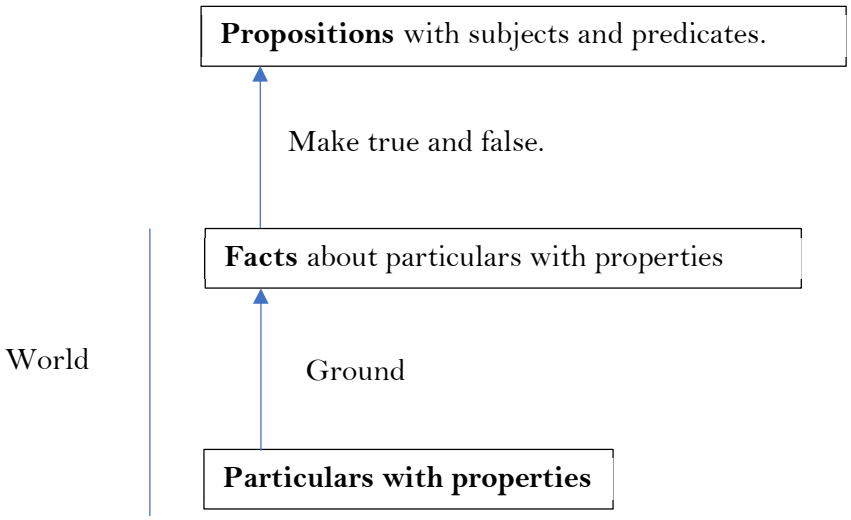
Why is this false? Well, because it is not a fact that I taught the class on April 20. We can schematically illustrate this as follows:



We also already know that particulars (concrete things in the real world) with their properties ground facts, but are not the same as facts:



Let's just say a tiny bit more about **propositions**. Russell says that a proposition "is a sentence in the indicative, a sentence asserting something, not questioning or commanding or wishing." Propositions have a subject and predicate(s), as Russell indicates on page 106, that is, propositions say *of something that it is a certain way*. For instance, the proposition 'Socrates is mortal' says of Socrates (subject) that he is mortal (predicate). Let's include this in our little picture.



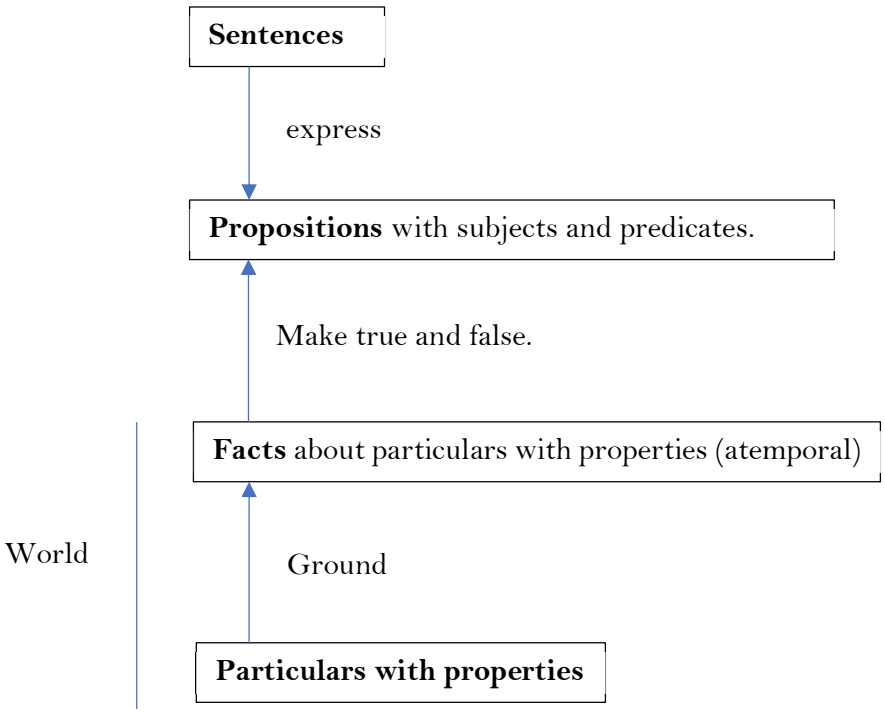
Now, I know that Russell says that propositions are sentences: "a proposition, one might say, is a sentence in the indicative." But this cannot be quite right. For consider the following two sentences.

"Socrates is mortal."

"苏格拉底是凡人."

These are different sentences, but they express the same proposition, that is, they mean the same thing. Russell does acknowledge this, maybe implicitly, towards the end of the paper. Finally, we can complete our illustration:

The combined view



Let's look at one example.

> The **sentence** "Julius taught the intro class on April 16, 2024"
>
> expresses the **proposition** ⌜Julius taught the intro class on April 16, 2024.⌝ This proposition predicates 'having taught intro on April 16, 2024' to the proposition's subject 'Julius.'
>
> This proposition is a true if, and only if, it is a **fact** that 'Julius taught the intro class on April 16, 2024.'
>
> And it is a fact if, and only if, Julius (the concrete person) did teach the intro class on April 16, 2024. That is, if there was such an **event**.

One interesting features of this fact-based metaphysics – the view that, fundamentally speaking, the world contains facts – is that the world has a lot of structure. Russell says:

> "Facts are, as I said last time, plainly something you have to take account of if you are going to give a complete account of the world. You cannot do that by merely enumerating the particular things that are in it: you must also mention the relations of these things, and their properties, and so forth, all of which are facts, so that facts certainly belong to an account of the objective world."

In a way, this might strike you as odd. Think about what the world is, all by itself, without anyone thinking about the world? You might think "well, it's just things in the void." Russell disagrees with this picture: the world in itself is rich in structure since it contains facts and not just objects. Let's consider some difficulties of this theory.

Existence. Now, you might think that this picture is a bit *too* convenient and that it leaves many questions open. Here are some of these questions. Consider the following proposition:

> ⌜Julius exists.⌝

If this is a true proposition, then there must be a fact that makes it true. But what property does it ascribe to me, Julius? It's hard to say/ Compare this proposition with this proposition:

> ⌜Julius teaches.⌝

'Teaching,' by contrast to existing, clearly is a property of Julius.

Negative facts. Finally, let's consider what Russell says about *false* propositions:

> "There are two different relations that a proposition may have to a fact: the one the relation that you may call being true to the fact, and the other being **false to the fact**." (105)

Consider a particular proposition:

'Socrates is alive.'

Following Russell's wording, this proposition *is false to the fact*. But false to what fact? Surely, there is no fact that makes it false, or is there? What about the following fact:

'Socrates is not alive'

This seems to mean:

It is not a fact that Socrates is alive.

But notice that it is not a fact that makes this sentence false. Rather, it is the absence of certain facts. But if we need such 'absences of facts' to account for false propositions, then it is strictly speaking not true that facts account for the truth and falsity of propositions.

Facts vs. names. Russell starts his paper arguing that the world contains facts and not just things. But maybe facts are just things: a complex kind of thing consisting of particular objects with certain properties. And propositions would, then, just be names for these facts. Russell disagrees with this. His reason is the following:

> Names without reference are meaningless, but statements (or propositions) without reference are not meaningless.

To see that names without reference are meaningless, consider a name that does not refer:

"Blicket"

Suppose "Blicket" refers to nothing and no one. Of course, we would say that, in this case, "Blicket" is just random noise. It's not name at all. By contrast, consider the statement:

"Socrates is dumb."

If Socrates is wise, then this sentence does not refer to any fact. It is, as Russell says, "false to the fact." But, importantly, it is not meaningless. We can perfectly well understand the sentence "Socrates is dumb" even if it is false. But we cannot understand the wannabe name "Blicket" if it refers to no one. Therefore, Russell says that *"propositions are not names for facts."* (105)