The knowledge argument and the zombie argument

I: the knowledge argument

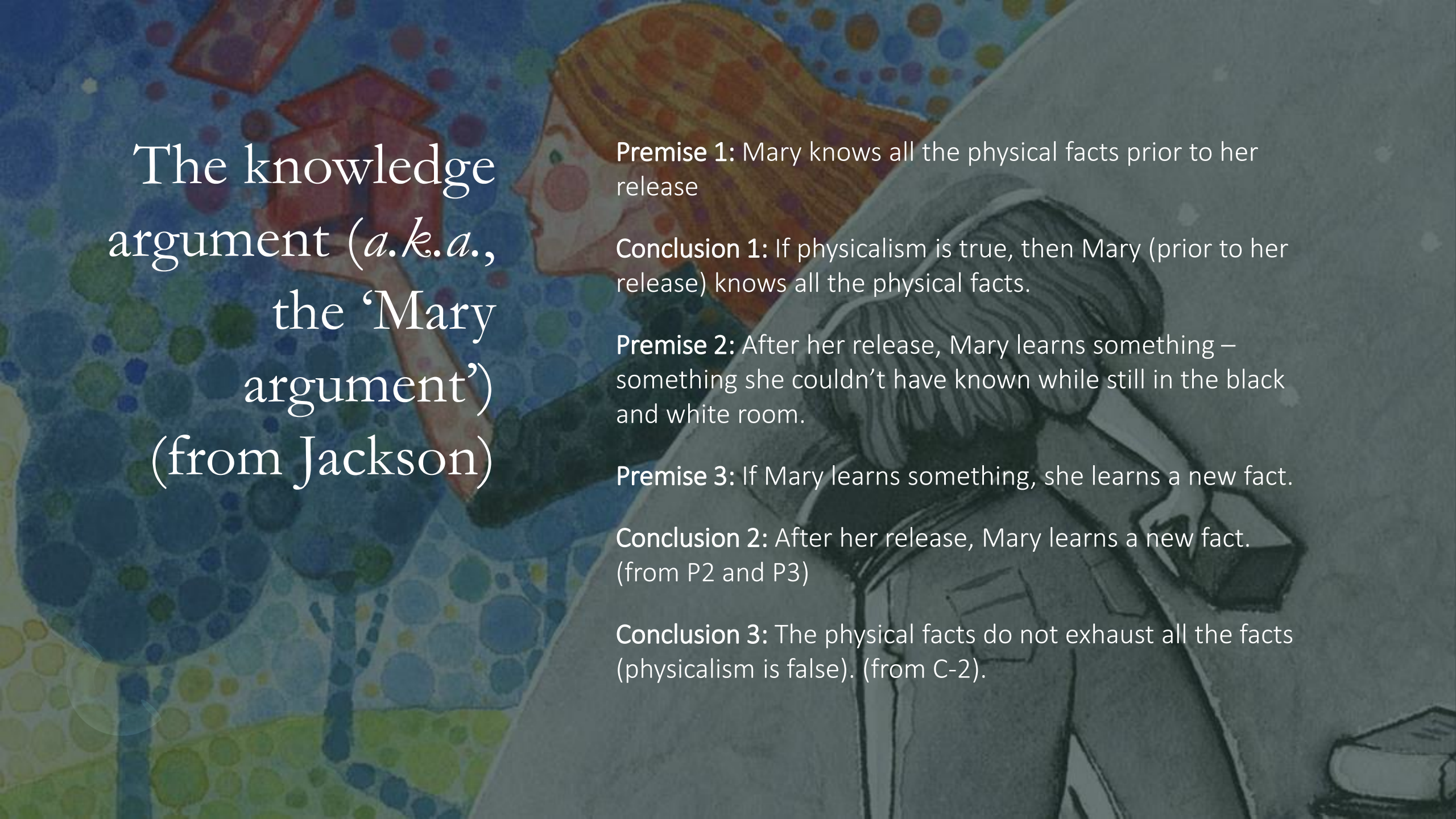
Set Up

Jackson's "Epiphenomenal Qualia" is the classic source for what is now called the 'knowledge argument' against physicalism.

- Jackson takes physicalism to be the thesis that "all (correct) information is physical information" (24). And he treats this as equivalent to saying that the facts – everything that is true of the actual world – are all physical.
- Physicalism, Jackson argues, is false: some (correct) information is nonphysical; put differently, some facts are nonphysical. (In terms of supervenience: the minimal physical facts do not determine all the facts.)

The knowledge argument

Jackson’s argument takes off from a thought experiment involving the brilliant super-scientist Mary, who has lived her life locked in a black-and-white room. (See p.130 for the details.) We can represent Jackson’s argument as follows:



The knowledge argument (*a.k.a.*, the 'Mary argument') (from Jackson)

Premise 1: Mary knows all the physical facts prior to her release

Conclusion 1: If physicalism is true, then Mary (prior to her release) knows all the physical facts.

Premise 2: After her release, Mary learns something – something she couldn't have known while still in the black and white room.

Premise 3: If Mary learns something, she learns a new fact.

Conclusion 2: After her release, Mary learns a new fact. (from P2 and P3)

Conclusion 3: The physical facts do not exhaust all the facts (physicalism is false). (from C-2).

The knowledge argument

Why does C3 (‘Physicalism is false’) follow from C1 (‘If physicalism is true, Mary (before her release) knows all the facts’) and C2 (‘Mary learns a new fact’)?

Answer: Granting physicalism, C1 entails that Mary knows all the facts before her release. But C2 says that Mary learns a fact upon release. So, if she really does learn a fact, it must be that she didn’t know all the facts while still in the black-and-white room. But she did know all the physical facts while in the room. So, there must be facts that aren’t physical facts, i.e., physicalism must be false.

- Why does C1 follow from P1? Answer: Actually, Jackson leaves it less than fully clear in “Epiphenomenal Qualia” why he thinks this. As we will see later, Jackson is assuming an understanding of physicalism that makes the inference from P1 to C1 valid. But that assumption turns out to be controversial (and not just for reasons having to do with consciousness and the mind-body problem).

The zombie argument

This argument, which (like Kripke's modal argument) traces back to Descartes's argument, has been developed systematically by David Chalmers. It aims to show that physicalism is false because of the conceivability of physical beings that are physically and functionally identical to phenomenally conscious beings but which fail to share those beings' phenomenally conscious states.

- Inverted you: Physically-functionally identical to you but phenomenally inverted. When you have a reddish visual experience, she has a greenish one, and vice versa …
- Zombie you: Physically-functionally identical to you but lacking phenomenal consciousness entirely. When you have a reddish visual experience, she has … nothing. There is "nothing it is like" to be them.

The zombie argument

The argument:

(P1) It is conceivable that there be zombies.

(P2) If it is conceivable that there be zombies, then it is metaphysically possible that there be zombies.

(P3) If it is metaphysically possible that there be zombies, then physicalism is false.

(C) Physicalism is false.

Comments and clarifications

On P1

- Assume: P is conceivable iff not-P is not a priori. Since it is not a priori that there not be zombies, it is conceivable that there be zombies.

On P2

- If zombies are conceivable, they are at least possible in the broad sense of metaphysical possibility.

On P3

- If zombies are metaphysically possible, then we know that the antecedent of the psychophysical conditional $P \rightarrow Q$ could hold even if the consequent does not. But that means that physicalism is false.

Jackson on Physicalism

To understand why Jackson thinks that the knowledge argument's P1 ("Mary knows all the physical facts prior to her release") entails C1 ("If physicalism is true, then Mary (prior to her release) knows all the physical facts"), we must turn to the "Postscript" and to Jackson's reflections on the necessary a posteriori.

Suppose someone were to argue like this:

- If physicalism is true, then a statement P describing all the actual physical facts necessitates a statement Q giving all the actual phenomenal facts. This follows from Supervenience Physicalism. Thus, the conditional P → Q is necessarily true. But it doesn't follow that the conditional is also *a priori*. And if it is not, then someone could know all the physical facts (like Mary) and yet not know all the phenomenal facts (also like Mary) even though physicalism is true. Compare: surely the physical facts necessitate the water facts. But not *a priori:* someone could know that there is H2O in her glass without knowing that there is water in her glass.

Jackson on Physicalism

This little speech might be bolstered by a remark like, "Didn't Kripke teach us that there can be necessary truths that are nevertheless *a posteriori*? If physicalism is true, then P → Q is just such a necessary a posteriori truth."

- Jackson believes that this argument rests on a mistake about how to interpret Kripke on the necessary a posteriori. (Note that there are physicalists who agree with him about this, so Jackson isn't obviously begging any questions here.)

Jackson on Physicalism

To review: Kripke argued that claims like this

(W) Water = H2O

are necessary a posteriori.

Jackson thinks that the correct way to understand W is as the result of a special sort of deduction.

Jackson on Physicalism

First, Jackson asks us to consider this claim:

> CLAIM: Competent users of "Water" understand (though they might not put it this way) that the term is a rigid designator whose reference is fixed by the description "the actual world stuff that plays the water role." [We might think of CLAIM as stating that the meaning of the word "Water" is something like the description "the actual stuff that plays the water role."]

But, Jackson continues, if CLAIM is true, then the following proposition is contingent a priori.

> (1) Water = the stuff that plays the water role.

(1) is *contingent* because water isn't what plays the water role in possible worlds where the superficially similar-yet-chemically different twin-water (XYZ) is the stuff that plays the water role. But (1) is *a priori* in that it follows from CLAIM, and so Jackson thinks we all know (1) solely in virtue of our competence with the word "water."

Jackson on Physicalism

Next, Jackson points out that empirical investigation reveals the following contingent a posteriori truth:

(2) The stuff that plays the water role = H2O.

But from (1) ("Water = the stuff that plays the water role"), (2) ("The stuff that plays the water role = H2O"), and the transitivity of numerical identity we can a priori deduce:

(W) Water = H2O.

Jackson on Physicalism

What this shows, Jackson suggests, is that if our concepts of natural kinds like water have the kind of structure that CLAIM attributes to them, we can deduce a priori that certain necessary truths like W hold. (Likewise, for necessary truths like "lightning is electrical discharge", "heat is mean molecular motion", "salt is sodium chloride", etc.).

The necessary a posteriori status of "water = H2O" is due to our being able to deduce it from "water = the watery stuff" (which we know a priori) and "the watery stuff = H2O" (which we know a posteriori). That is why Jackson thinks that with enough physical information someone with the right concepts and deductive abilities ought to be able to know all the water facts.

Jackson on Physicalism

How does this relate to Mary and phenomenal consciousness? Well, Jackson believes that if physicalism is true, a priori physicalism must be true. It isn't just that P → Q is necessarily true; it is also *a priori* in the sense that if the physical facts are the only facts there are, then any being who knows P will be able to perform a deduction to Q like the one that got us from (1) and (2) to W. Since Mary can't do this, despite her awesome physical knowledge and powers of deduction, then physicalism is false.

- What motivates the thought that physicalism requires a priori physicalism is the idea that, without the possibility of *a priori* deduction, claims like W and P → Q would be *epistemically brute*.

Jackson on Physicalism

So, Jackson believes we all share the intuition that when Mary has her first chromatic colour experience, she learns something new (e.g., what it's like to experience red). And Jackson also believes that this intuition supports an important epistemic premise:

> The Non-Deducibility Claim (NDC): There are truths about phenomenal consciousness that cannot be a priori deduced from the totality of physical truths.

Whereas someone with the right concepts, physical information, and deductive capacities could a priori deduce (and thus know) all the facts about water on the basis of facts about H2O, the knowledge argument shows that the phenomenal facts *cannot* be a priori deduced from complete physical information. (The conceivability of zombies demonstrates the same).

A priori vs. a posteriori physicalism

One physicalist response to the anti-physicalist arguments is to reject NDC – i.e., to assert that the phenomenal facts can be deduced a priori from the physical facts. This is the strategy of:

> A priori physicalism (a.k.a. 'Type-A materialism'): Supervenience physicalism (as we have understood it) plus the claim that all truths can in principle be a priori deduced from the physical truths.

- agrees with anti-physicalists like Jackson and Chalmers about how to *interpret* physicalism, but:
- disagrees with anti-physicalists about the truth of NDC. Hence, the physical-phenomenal conditional, $P \rightarrow Q$, is not only necessarily true but knowable a priori for anyone with the requisite concepts.
- entails that all concepts (including phenomenal ones) are analyzable in terms of structure and function. [Often goes with a neo-Frege-Russell view.]
- Proponents: Daniel Dennett; David Lewis; Keith Frankish (?)

A priori vs. a posteriori physicalism

The main alternative:

A posteriori physicalism (a.k.a., 'Type-B materialism'): Minimal physicalism plus the denial of the claim that all truths can in principle be a priori deduced from the physical truths.

- disagrees with anti-physicalists like Jackson and Chalmers about how to *interpret* physicalism, but:
- agrees with anti-physicalists about the truth value of NDC. The physical-phenomenal conditional, $P \rightarrow Q$, is necessarily true but not knowable a priori for anyone with the requisite concepts. (That is, the physical-phenomenal conditional $P \rightarrow Q$ is necessarily true but only a posteriori.)
- typically goes along with the denial that all concepts are analyzable in terms of structure and function. (At the heart of the so-called "Phenomenal Concepts Strategy." [Usually goes with a rejection of neo-Frege-Russell views.])
- Proponents: Brian Loar, Ned Block, David Papineau, Janet Levin (and many others)