**The answers provided below are sometimes longer and more thorough than necessary.

1. Answer: According to philosophers who use the notion of metaphysical supervenience to define physicalism, physicalism is true at a possible world, w, iff any world which is a minimal physical duplicate of w is a duplicate of w *simpliciter* (i.e., a complete duplicate). Put differently, the physical properties/events/facts metaphysically necessitate (i.e., are metaphysically sufficient for) all other properties/events/facts about the world, including all the mental ones. It follows that, corresponding to every mental difference between two possible worlds, there will be some physical difference. Like a pointillist painting is 'nothing over and above' a complex assortment of coloured points, mental properties/events/facts are nothing over and above complex aggregates of physical properties/events/facts.

[To receive full marks, the answer must state that, according to supervenience physicalism, two possible worlds that are minimal physical duplicates are thereby complete duplicates. (You can award a bonus mark if they explain what makes a physical duplicate "minimal", but the main point should be on the idea of duplication). Give partial marks if: (i) the answer doesn't successfully explain supervenience physicalism but does mention possible worlds, or (ii) the answer fails to discuss possible worlds but communicates the intuitive idea that the mental is somehow 'nothing over and above' the physical or of 'coming along for the ride' with the physical or of being a higher level 'pattern' of physical phenomena. It is incorrect to answer that mental entities are physical entities.]

2.

(a) the completeness question (what does it mean to say that *everything* is physical?) -- correct

(b) the condition question (what does it mean to say that everything is *physical*?) – incorrect

(c) the truth question (is physicalism *true*?) -- incorrect

3. The difference between reductive and non-reductive physicalism is that reductive physicalists hold that mental properties are *strictly identical to* physical properties (e.g., that the experience of pain just is a certain type of brain state), whereas non-reductive physicalists hold that mental and physical properties are distinct – i.e., they uphold property non-identity.

[Full marks for isolating the question of property identity/non-identity as the relevant difference.]

4. Non-reductive physicalists are usually motivated by the wish to accommodate the "multiple realizability" of mental states. Mental states are multiply realizable if two beings could share the same mental states despite differing physically (e.g., anatomically). To accommodate this possibility, one must deny that mental properties are identical to physical properties, since if a mental property were identical to a certain physical property, then the physical property in question would be necessary for the instantiation of the mental state, which contradicts the supposition of multiple realizability. To uphold physicalism, they can assert that mental states are nevertheless metaphysically necessitated/determined by physical properties. On the usual way of understanding this situation, mental state terms (e.g., 'pain' or 'belief') pick out second-order or functional role properties that are "topic-neutral" (or noncommittal) with respect to the underlying nature of the state realizing the functional role. If physicalism is true, then these second order properties are "realized" or "implemented" by physical properties (the actual physical states that play or "fill" the causal role in question).

[1 mark for explaining the notion of multiple realizability in terms of sameness of mental property despite variation in physical property. 1 mark for explaining why multiple realizability is inconsistent with reductive physicalism/type-identity theory. 1 mark for noting that nonreductive physicalists replacing identity with a different relation of metaphysical necessitation, such as "realization". Give partial marks if the student uses the phrase 'multiple realizability' without explaining it fully.]

5. Supervenience physicalism is consistent with *both* reductive and non-reductive physicalism. In particular, reductive and non-reductive physicalists agree that the mental metaphysically supervenes on the physical (alternatively: that the physical metaphysically *necessitates* the mental) but disagree about the specific non-modal

relation in virtue of which mental properties metaphysically supervene on physical properties. Reductive physicalists claim that this relation is one of strict numerical identity: that mental properties are one and the same as a particular type of physical property. By contrast, nonreductive physicalists posit a relation that is strong enough to allow metaphysical supervenience/necessitation but weak enough to permit multiple realizability (e.g., 'realization', 'constitution', or 'determination').

[1 mark for noting that supervenience physicalism is silent on the reductive/nonreductive debate. 1 mark for explaining that the difference between the two views lies in how they explain the metaphysical necessitation of the mental on the physical]

6. One proposal (from Yablo) is that the metaphysical relation in virtue of which the mental metaphysically supervenes on the physical is the determinable-determinate relation. On this view, a mental property like pain is related to its various physical realizers (e.g., C-fibres, silicon chips, etc.) in the same way that a colour (e.g., red) is related to one of its determinate shades (e.g., crimson or scarlet). Another proposal we considered (which might be seen as a development of the determinable-determinate proposal) is that the relation in virtue of which the mental metaphysically supervenes on the physical is the relation of proper parthood (e.g., Wilson). On this latter view, the causal powers of a mental state are a *proper subset* of the causal powers of its physical realizer. This is the 'causal subset strategy'.

[1 mark for identifying either Yablo or Wilson's views as attempts to fill in supervenience physicalism with a more precise account of the mental-physical relation. 1 mark for explaining the view correctly.]

7. There are various metaphysical possibilities that nonreductive physicalism prohibits but that property dualism allows. The most obvious is a zombie world: a possible world that is a minimal physical duplicate of our actual world but in which no phenomenal properties at all are instantiated. Another example is a 'qualia inversion' scenario, in which two possible worlds are physically exactly similar but are phenomenally inverted. Both zombie and qualia inversion scenarios violate the assumption of metaphysical supervenience, since they allow mental difference without any corresponding physical differences. Accordingly, the nonreductive physicalist must deny that these are genuine possibilities, whereas the property dualist can allow for them.

[1 mark for describing a scenario in which a mental feature fails to metaphysically supervene on physical features. 1 mark for explaining why this is consistent with property dualism but inconsistent with nonreductive physicalism.]

8. 'Hempel's Dilemma' is a challenge to physicalism that arises when we attempt to specify what we mean by 'physical'. In particular, either 'the physical' refers to the postulates of current physical theory or to the posits of some future as-yet unspecified physical theory. If the former, then physicalism is almost certainly false, since contemporary physics is very likely either incomplete or partly mistaken. If the latter, then we don't actually understand what physicalism says, and so the proposal is too vague to be of use. (It also runs the risk of making physicalism true on the cheap and so uninteresting if the category of the 'physical' is so open-ended that it refers to whatever entities figure in a true theory of reality). On the informative reading, it is false. On the true reading, it is uninformative.

[1 mark for noting that the challenge to physicalism is a dilemma that arises when we attempt to specify what we mean by 'physical'. 1 mark for correctly explaining the first horn. 1 mark for correctly explaining the second horn.]

9. No. According to Montero, although debates about physicalism should be dropped, many of the central issues that have been debated under the moniker of physicalism can be recast as questions about the fundamentality or non-fundamentality of the mental. The position traditionally called 'physicalism' can be reformulated as the position that the mental is a nonfundamental feature of reality.

[1 mark for noting that Montero's answer is 'no'. 1 mark for explaining that the debate can be recast in terms of the (non)fundamentality of the mental.]

Section 2: mental causation

1. Papineau's rests his causal argument for physicalism on three plausible-seeming claims and one implicit premise:

Efficacy: Mental events sometimes sufficiently cause physical events (and do so in virtue of their mental properties.)

The Causal Completeness of the Physical (a.k.a. Causal Closure): Every physical event has a fully sufficient physical cause.

No Systematic Overdetermination: The physical effects of mental causes are not systematically causally overdetermined.

Exclusion: No effect has more than one sufficient cause unless it is causally overdetermined. (implicit premise)

More specifically, if we want to uphold the causal sufficiency of a mental event for bodily movement (Efficacy), the causal sufficiency of a physical (e.g., neural) event for bodily movement (Completeness), and the absence of systematic overdetermination (No systematic overdetermination and Exclusion), then we will be forced to conclude that mental properties are not distinct from physical properties, and so that physicalism is true.

[1 mark for each correctly identifying each assumption; 1 mark for explaining how they combine to yield a physicalist conclusion]

2. We can motivate each of Papineau's premises as follows:

- Rejecting Efficacy involves rejecting (the widely accepted) 'Alexander's dictum': that 'to be is to have causal powers'. It might also seem to imply that the experience of intentional agency is illusory, and would make introspective knowledge of our mental states mysterious (on the assumption that knowledge of x entails bearing some causal connection to x).

- Were we to reject Completeness, we would seem to be contradicting central claims of contemporary physics. (Most often, the contradicted claim is taken to be a conservation law of physics, though Papineau is actually more careful than this).

- Were we to reject No Systematic Overdetermination, we would seem to be treating every case of mental-to-physical causation on the model of death by firing-squad: i.e., of a person who is shot to death by multiple simultaneous bullets to the chest each of which is sufficient to cause his death. Overdetermination cases are possible, but the idea that they occur *systematically* and *pervasively* (e.g., every time you intentionally act) seems incredibly implausible.

- Were we to reject Exclusion, we would be denying that an event's having more than one sufficient cause is, by itself, sufficient for its being causally overdetermined. But it is not immediately obvious what such a case could be.

[1 mark for one correct motivation for each premise. The answer does not need to mention all motivations.]

3. Kim argues that the causal argument – in Kim's terminology the causal exclusion argument – is not an argument for physicalism but for *reductive* physicalism. This is because, among physicalists, only reductive physicalists hold that mental properties are not distinct from physical properties – i.e., that they are identical. Hence, if the causal argument succeeds, as Kim believes it does, it refutes not only dualism but also nonreductive physicalism.

[1 mark for noting that the causal argument seems to support reductive physicalism in particular. 1 mark for correctly explaining why this is so.]

4. According to Bennett, nonreductive physicalists (but not dualists) are free to reject the implicit assumption we have called "Exclusion": that the only case in which a single event has two distinct fully sufficient causes is a case of causal overdetermination (of the firing squad variety).

[1 mark for successfully identifying Exclusion as the assumption that Bennett believes nonreductive physicalists are free to reject]

5. The counterfactual that is either false or vacuously true if nonreductive physicalism is correct is (O2): that had the same physical (e.g., neural) cause still occurred but the mental cause not occurred, the same behavioural effects would still have occurred.

6. Bennett claims that most dualists will regard (O2) as non-vacuously true. Their (alleged) commitment to Causal Completeness of the Physical will lead them to regard the physical cause as fully sufficient for the behavioural effect. Meanwhile, their denial of supervenience physicalism will lead them to regard the occurrence of the physical cause in the absence of the mental cause as a genuine metaphysical possibility. Because the antecedent of (O2) is not metaphysically impossible, the truth of the counterfactual is not only true but non-vacuously true.

[1 mark for noting that, according to Bennett, dualists will wish to uphold (O2) as non-vacuously true. 1 mark for noting how Causal Completeness leads them to regard the conditional as a whole as true. 1 mark for noting how their denial of supervenience physicalism leads them to regard the truth of the conditional as non-vacuous]

7. Gibb's version of interactionist dualism is non-standard in how it conceptualizes mental-to-physical causation. Specifically, whereas standard versions of interactionism (e.g., Cartesian dualism) depict the causality involved in mental causation in terms of a form of energy transfer (e.g., as the triggering of a physical event by some mental event), Gibb's dualism rejects this assumption. On Gibb's version of dualism, the causality at work in mental causation is that of double prevention. On this view, mental events do not directly cause physical events but, instead, enable a certain process of physical causation to occur.

Section 3: Naturalization intentionality/mental representation

1. Intentionality refers to the relation of 'aboutness', 'of-ness', or 'directedness' that obtains between a mental state and its 'object' or subject matter. In the case of descriptive mental states, this can also be understood as the 'semantic' or 'representational content' of the mental state. Intuitively, the content of a mental representation is what the representation 'says' about the world. It is the 'condition' that that reality must satisfy in order for the representation to count as true, accurate, or successful.

2. Compared with familiar physical relations, the intentional relation between representation and represented — 'aboutness' or 'of-ness' – is puzzling because of the apparent lack of causal connection between the two elements of the relation. This is most salient in the case in which what is represented is fictional or nonexistent – e.g., Santa, Pegasus or the Fountain of Youth. Relatedly, we routinely represent not only how things are but how thing might have been different – e.g., we believe and entertain various counterfactual conditionals. But, as a general metaphysical matter, the instantiation of a relation implies the existence of each relatum. If so, then the puzzle is to explain how we become related in thought to nonexistent objects and states of affairs (e.g., fictional entities and counterfactual situations). Even if we set aside such entities, we regularly think about concrete entities that seem to be causally 'absent' in the sense of not exerting any direct causal influence on us: e.g., objects and events that are remote from us in time and space. On the face of it, there are no causal constraints on which beings we can enter into intentional relation with – e.g., there is no limit on the amount of time that it takes to refer to, or make representational 'contact' with, a far-away object, nor is there any obviously detectable physical process which mediates that relation. And (this is the most puzzling part) the intentional relations we bear to distal phenomena seems nevertheless to causally impact our behaviour: the contents of our thought determine what we say and do. If so, then we will not have understood a substantial part of our human behaviour until we understand the relation of aboutness. Familiar mechanical principles seem to offer no insight into how a physical system like the human brain/body – which presumably operates according to 'local' mechanical principles (collisions, transferences of energy, etc.) –

coordinates its behaviour with entities which are not causally affecting it. In contrast to reflex behaviours, representation-mediated behaviour looks like a weird form of action at a distance.

[1 mark for noting that we seem to be able to represent things – and so be intentionally 'related' to – that do not exist and/or things that exist but that aren't causally affecting us. 1 mark for giving instructive examples (e.g., entities that are fictional, counterfactual, abstract, past or future, and too far away to have effects on us). Bonus mark for noting how these puzzles extend to human behaviour insofar as behaviour is mediated by representational activity.]

3. X reliably indicates Y if X item carries information about Y in virtue of systematically correlating with Y. Many theorists have believed that reliable indication will figure centrally in a naturalistic account of intentionality/mental representation. This is partly inspired by the observation that reliable indication seems to be sufficient for what Grice called "natural" meaning, or what we could also call "natural signs" or "natural representations". Examples include:

- smoke reliably indicates fire; thus, smoke *means* fire.
- 47 tree rings reliably indicates 47 years; thus, 47 tree rings *represent* how old the tree is.
- Spots reliably indicate measles; spots are a *sign* of measles.

The main contrast with natural meaning/signs/representations is what Grice called "conventional" meaning (e.g., the way in which a red traffic light or a 'stop' sign means *stop*). With conventional meaning, the relationship between the representation and represented is arbitrary and a matter of social convention: participants in the linguistic community collectively *decide* to treat one thing as a sign of something else.

The attraction of starting with natural meaning rather than conventional meaning is that natural meaning is an unmysterious and familiar aspect of nature: it is totally ubiquitous. By contrast, conventional meaning depends on linguistic intentions, which are a type of mental representation. Further, there are rudimentary kinds of mental representation which seem like they *might* be a matter of natural meaning, notably sensory representation and/or sensory recognitional concepts. Phenomenologically, seeing something seems to involve a kind of perceptual contact with it. And sub-personally, neuroscientists frequently speak of specialized 'feature detectors' that reliably respond with heightened activation in the presence of certain types of stimuli. If all other mental representations depend somehow on sensory representations, and if sensory representation represent what they do by detecting/reliably indicating the presence of certain environmental stimuli, then an account of mental representation in terms of reliable indication seems like a promising place to start our theorizing.

[1 mark for noting correctly explaining the notion of reliable indication. 1 mark for noting its connection with 'natural' as opposed to 'conventional' meaning, the former of which is an unmysterious and unproblematically physical phenomenon. 1 bonus mark if the answer points out that sensory representation, in particular, may be a matter of detecting (/reliably indicating) various sorts of environmental information, and that, given its relatively basic status, sensory representation is a natural place to start theorizing representation.]

4. These problems arise for the claim that X represents Y iff X reliably indicates Y. In each case, the problem arises from the fact that reliable indication (unlike representation) metaphysically necessitates the presence of whatever it indicates.

The misrepresentation problem is that if reliable indication is <u>necessary</u> for representation, then misrepresentation – specifically, the tokening of a representation-type in the absence of what it represents – is impossible. If something is not causally present, then it cannot be indicated, and if it is not able to be indicated, then (according to the proposed analysis) it cannot be represented.

The disjunction problem is that if reliable indication is <u>sufficient</u> for representation, then whatever is reliably indicated by a representation-token in a given type of circumstance is thereby represented by it. This will mean that intuitively far too many situations will be included in the content of the representation. E.g., what

we might have thought was a representation of sheep turns out to be a representation of a sheep *or* a goat on dark, foggy night (*or* whatever else might be reliably trigger the representation-type to be tokened in a given sort of circumstance). Hence, representations will have widely disjunctive content.

Apart from the fact that we want a theory of representation to allow for the possibility of misrepresentation and for determinate (rather than widely disjunctive) content, arguably the deeper issue these problems bring to our attention is that we have not yet been shown how, through reliable indication, we can even start to understand the most important and puzzling feature of representation, which is how the internal state of a physical system is able to transcend what is causally local (e.g., what is causally indicated) and represent something in its absence. This is something we find in an elementary way whenever a representation misrepresents. To be representationally related to the world, it seems that a physical system's internal states and processes must possess a certain sort of independence from what is causally present to it: it must be able to represent something even when the represented thing is absent.

[1 mark for noting that the problem arises from the putative necessity (in the case of the misrepresentation problem) or putative sufficiency (in the case of the disjunction problem) of reliable indication for mental representation. 1 mark for explaining the undesirable consequence in question (whether 'ruling out' the possibility of misrepresentation, or 'ruling in' every possible trigger into the representation's content). 1 mark for correctly stating the explanatory constraint that the discussed objection introduces for a theory of representation – namely, to avoid the undesirable consequence that they have just discussed.]

5. One such proposal is indicator teleosemantics, according to which X represents Y if and only if X has the proper function of reliably indicating Y. Indicator teleosemantics deals with the disjunction problem by restricting the content of a representation-type to the class of entities the presence of which it has the *proper function* of detecting. What makes a representation about a certain type of entity (rather than the wide disjunction of entities that will ever cause it) is that the representation has the proper function of indicating the presence of entity in question (and not all the other things that might trigger it).

Consumer-based teleosemantics is another teleosemantic proposal. I did not define this view in terms of necessary and sufficient conditions. But the core idea is that what a representation represents is determined by the representation's causal contribution to a larger evolutionarily selected, life-sustaining process, rather than simply by what it serves to indicate. In the case of the frog and the fly, for example, the visual representation has the content that it does not in virtue of being reliably caused by certain type of stimuli, but in virtue of the use to which the representation is subsequently put – e.g., in feeding. The content of the visual representation is fixed not by causal inputs but by causal outputs. (Saying more about how exactly this story goes requires some care. It may be that the content of the visual state has both an 'indicative' and a 'directive' aspect; e.g., as food source (descriptive) that is to-be-eaten (directive). So, this view does not require that there is no indicative element to the content, but whatever indicative content it has will be fixed in relation to the larger routine within which the representation is functionally embedded: so, the object is represented in terms of the actions that can be performed on it.)

Indicator and consumer-based teleosemantic accounts deal with the misrepresentation and disjunction problems in a similar way – by narrowing the content of the representation down to whatever it has the function of representing. They will handle the misrepresentation problem by allowing that, as is generally true of systems with proper functions, it is possible for representational systems to malfunction. In the case of indicator teleosemantics, misrepresentation consists in firing in the presence of something other than what it has the function of reliably indicating. (Even when reliably triggered by something other than what it represents, the representation will nevertheless represent whatever is triggering it as being the type of entity that it has the proper function of indicating.) In the case of consumer-based teleosemantics, misrepresentation can be spelled out in terms of whether the representational token results in the behavioural effects for which the representation was selected. Parallel moves will be pursued in response to the disjunction problem: in each case, though the system is causally sensitive to disjunction, this is not what the representation was selected to be sensitive to.

[1 mark for correctly explaining the notion of proper function. 1 mark for explaining how it can be invoked to help explain content determination for mental representations – whether in an indicator or consumer-

based fashion. 1 mark noting how proper function can be used to narrow down a representation's content from all the information that it is causally sensitive to (thereby solving the disjunction problem) or to explain the possibility of misrepresentation as a special case of malfunction (thereby solving the misrepresentation problem).]

5. A problem for indicator teleosemantics is the functional indeterminacy problem, which is the problem that teleofunctions alone deliver highly indeterminate contents, since evolution is insensitive to distinctions we might want to draw concerning the content of the state (e.g., whether the frog represents a fly, a small black moving thing, or even a certain optical pattern that does not distinguish the frog from its surroundings).

A problem for the teleosemantic approach more generally comes from Swampman: a being who popped into existence would lack any proper functions according to the etiological account of proper functions that teleosemantic theories have tended to assume. (According to these accounts, the proper function of some trait is whichever subset of its causal effects contributed to its being evolutionarily selected and explains its persistence within the population today). Swampman has no evolutionary history, so it follows on teleosemantic views that assume an etiological account of proper function that his mental states have no representational content. Many find these verdicts about the mental life of Swampman counterintuitive.

One might object that the reductive, mechanistic approach is misguided on some other ground. Perhaps it does not capture what we really want from a theory of intentionality, such as the way it manifests phenomenologically for the subject (since we've said nothing of subjective consciousness in these accounts, but only of seemingly sub-personal phenomena like informational links and evolutionary histories). Or perhaps one might object that it does not seem like it will generalize to all the intentional phenomena we ultimately want to have explained (we are still just at the level of perceptual representations). Or one will object that these accounts do not give adequate weight to the normativity of representation and/or the constitutive link between representation and objectivity. (While we did not discuss this, many theorists will dispute whether mere biological teleology is sufficient for normativity, and likewise many theorists will dispute whether any of the materials available to a broadly causal approach actually deliver a notion of objective representation as opposed to mere causal sensitivity. Those who have such misgivings often give significant emphasis to human social (e.g., linguistic) practices in accounting for intentionality, and this will typically fit uneasily with any naturalistic project.)

[1 mark for a coherent objection, and 1 mark for identifying which commitment the objection targets]

Section 4

1.

mind

body

possible

possible

mind

body

mind

body

2. Smart's objection was that C1 – that it is possible for my mind to exist without my body existing – does not offer support to C2 – that my mind is not my body. This is because, according to Smart, minds (or mental states) could be contingently identical to brains (or brain states). In Smart's view, contingently true identity statements are commonplace in science. Although it is perfectly true that water is identical to H2O, Smart claims, it could have been something else: empirical investigation might have revealed it to be XYZ. Likewise,

Smart reasons, neuroscience has (or eventually will) reveal pain and CFF to be identical, even though there is a possible world in which pain is identical to something else.

[1 mark for noting that Smart's objection attacks the transition from C1 to C2. 1 mark for noting that the relation between mind and body could be one of contingent identity. 1 mark for noting that contingent identities appear to be commonplace in science, as when one makes the empirical discovery that water is H2O, even though it could have been something else.]

3. Smart's argument assumes that natural kind terms like 'water' and phenomenal terms like 'pain' are 'non-rigid designators', i.e., that in different possible worlds, they refer to different things. A clear example of non-rigid designation is a definite description like 'the prime minister of Canada'. In the actual world, this picks out JT (since JT is who happens to satisfy the description). But there are possible worlds in which JT exists but isn't the PM. (Being PM of Canada is an accidental rather than essential property of JT, and the proposition that JT is PM is contingently rather than necessarily true). By contrast, Kripke argues that natural kind terms like 'water' and 'H2O' as well (and phenomenal terms like 'pain') are rigid designators: in every possible world in which the terms refer, they refer to the same entities. If so, then identity statements like 'water is H2O' and 'pain is CFF' are, if true, necessarily true (since, in virtue of being rigid, the two terms of the identity pick out the same objects in all possible worlds). If Kripke is right about this, then none of the theoretical identity statements that Smart offers are contingent, and Smart's strategy for replying to Descartes fails. If Descartes is right that mental states could exist apart from physical states, then they are, indeed, distinct.

[1 mark for noting that explaining that Smart seems to falsely assume that natural kind terms and phenomenal terms non-rigidly designate their referents. 1 mark for explaining nonrigid designation with the example of a definite description like 'the F'. 1 mark for noting that if natural kind and phenomenal terms are rigid designators, as Kripke argues they are, then Smart's strategy fails, since in that case theoretical identities are necessarily rather than contingently true.]

4. Kripke's response is to explain away (i.e., to reject) the apparent contingency of identities like "water is H2O". by disputing that we are actually imaging what we take ourselves to be. When we think we're imagining a possible world in which water is not H2O but something else (e.g., XYZ), what we're *really* imagining is a possible world in which something with different essential properties than H2O has some of the same observable *effects* that H2O has in the actual world. Mores specifically, we are imagining a possible world in which something other than H2O has the same *perceptual appearances* in us that H2O has in the actual world. Once we see this, we also see why our initial description of the case depended on a confusion. We were confusing the inconceivable proposition (that water could exist without H2O) with a different, conceivable proposition (that something other than H2O could have the same sensory effects on us – could have been causally responsible for the same watery appearances – that H2O has on us). If so, then the scenario we are imagining gives us no reason to deny that water and H2O are necessarily identical.

[1 mark for noting that Kripke 'explains away', or rejects, the appearance of contingency. 1 mark for noting that Kripke thinks we are confused about what we are imagining when we think we are imagining water as something other than H2O. 1 mark for noting that what we are really imaging is something other than H2O having the same sensory effects that H2O has in the actual world (or, more simply, that we are confusing water itself with the appearance of water.]

5. According to Kripke, no. While we can easily imagine something with different essential properties than water sharing the same sensory appearance that water actually has, we cannot similarly imagine something with different essential properties than pain sharing the same sensory appearance that pain actually has. This is because the appearance of pain is an essential property of pain: anything that has the appearance of pain just is pain. To make 'pain' analogous to 'water', we would need to be able to say: "it could have been that something with different essential features than pain felt the same way pain feels". But that is false: whatever feels like pain is pain. So, unlike familiar examples of theoretical identification in science, the apparent contingency of a statement like 'pain is CFF' cannot be explained away. In this case, the appearance of contingency really does provide reason to reject the alleged psychophysical identity (as well as any other a posteriori necessity between a physical and a phenomenal property).

[1 mark for noting that Kripke's answer is 'no'. 1 mark for noting that the question, according to Kripke, is whether we can imagine something with different essential properties than pain sharing the same appearance that pain has. 1 mark for noting that, according to Kripke, this is *not* something we can imagine. 1 mark for his proposed reason: namely, that if we imagine something with the appearance of pain, then we are imaging pain itself (since pain is an appearance or feeling).]

6. Jackson's argument rests on a brilliant super-scientist, Mary, who has lived her life locked in a black-and-white room and is one day released from her room and has her first chromatic colour experience. We can represent Jackson's argument as follows:

(P1) Mary (prior to her release) knows all the physical facts.

(C1) If physicalism is true, Mary (prior to her release) knows all the facts.

(P2) After her release, Mary learns something (something she couldn't have known while still in the black-and-white room).

(P3) If Mary learns something, she learns a new fact.

(C2) Mary learns a fact.

(C3) Physicalism is false.

7. Jackson defends epiphenomenalism on the basis of this argument.

8. The two main physicalist responses are a priori/Type-A physicalism and a posteriori/Type-B physicalism. The former attacks the premise that there is an epistemic/explanatory gap between physical and phenomenal concepts, while the latter attack the premise that if there is an epistemic/explanatory gap then there is a metaphysical gap (or failure of metaphysical supervenience).

How Type-A physicalists treat the different arguments. In response to Descartes'/Kripke's and Chalmers' conceivability arguments, Type-A physicalists will attack the conceivability premise (e.g., that one can genuinely conceive of one's phenomenal state in the absence of any physical state, or that one can conceive of one's physical states in the absence of any phenomenal states). In the case of the knowledge argument, most Type-A physicalists rejects Premise 2 – the claim that Mary learns something new when she leaves the room. (The main exception, here, is David Lewis, who is clearly a Type-A/a priori physicalist but tries to accommodate Premise 2. However, the sense in which Lewis grants Premise 2 is very different than how either anti-physicalists and Type-B/a posteriori physicalists do. According to Lewis, Mary does learn something when she leaves her room, but what she learns is not a new proposition (as both anti-physicalists and Type-B physicalists will typically claim). Rather, she acquires new abilities: e.g., to recognize coloured objects, sort them into categories, imagine them, etc. Notably, there is no new information or proposition that Mary comes to grasp when she has her first chromatic colour experience. She merely has new functionally specifiable dispositions.)

How Type-B/a posteriori physicalists respond to the anti-physicalist arguments. They reject the second premise of both Descartes'/Kripke's and Chalmers' arguments (the assumption that conceivability entails metaphysical possibility). In the knowledge argument, they can be seen as rejecting the inference from P1 to C1: that if Mary knows all the physical facts before her release then, if physicalism is true, she must know (or be able to a priori deduce) all the facts. Type-B/a posteriori physicalists allow that there are certain facts about colour experience that Mary is not in a position to know until she can think about colour experience introspectively using a phenomenal concept. (An alternative way to develop Type-B/a posteriori physicalism is to grant that C1 is true but reject P3: i.e., to assert that Mary already knows all the facts from inside the room and that what happens when she leaves the room is she now thinks about "an old fact in a new way" (i.e., under a different mode of presentation). This assumes a coarser-grained metaphysics of facts, but the guiding idea behind Type-B physicalism remains the same: grant to the anti-physicalist that there are no a priori connections between Mary's physical concepts and her phenomenal concepts of colour experience, but

then insist that the relationship between physical properties and phenomenal properties may still be metaphysically necessary. Metaphysical necessities need not (and often are not) deducible a priori. More often, they will claim (following Smart's lead) they are the result of a (non-deductive) inference to the best explanation.

[1 mark for noting that the difference is between rejecting the epistemic premise and rejecting the metaphysical premise. 1 mark for naming the type of physicalism that this corresponds to (Type-A or Type-B). 1 mark for noting that the difference concerns whether metaphysical necessities need to be a priori. Grant a bonus mark for a thorough explanation or an illustration in terms of one of the arguments.]

One reason to prefer Type-B physicalism is that it can be concessive to anti-physicalists intuitions. As far as introspective evidence is concerned, they can grant nearly anything the anti-physicalist wishes to say about the irreducible differences between how experiences seem first-personally and how they seem third-personally, while attributing these irreducible differences to our distinctive ways of representing experiences. If this can be made to work, then it will have the virtue of taking the introspective evidence as seriously as the anti-physicalist does without ceding any metaphysical ground to the anti-physicalist. By contrast, Type-A physicalism courts the reply that they are simply denying the obvious or are failing to grasp what makes experience distinctive among the elements of reality. The debate between Type-A physicalists and anti-physicalists is apt to result in a sense that the two sides failing to understand each another.