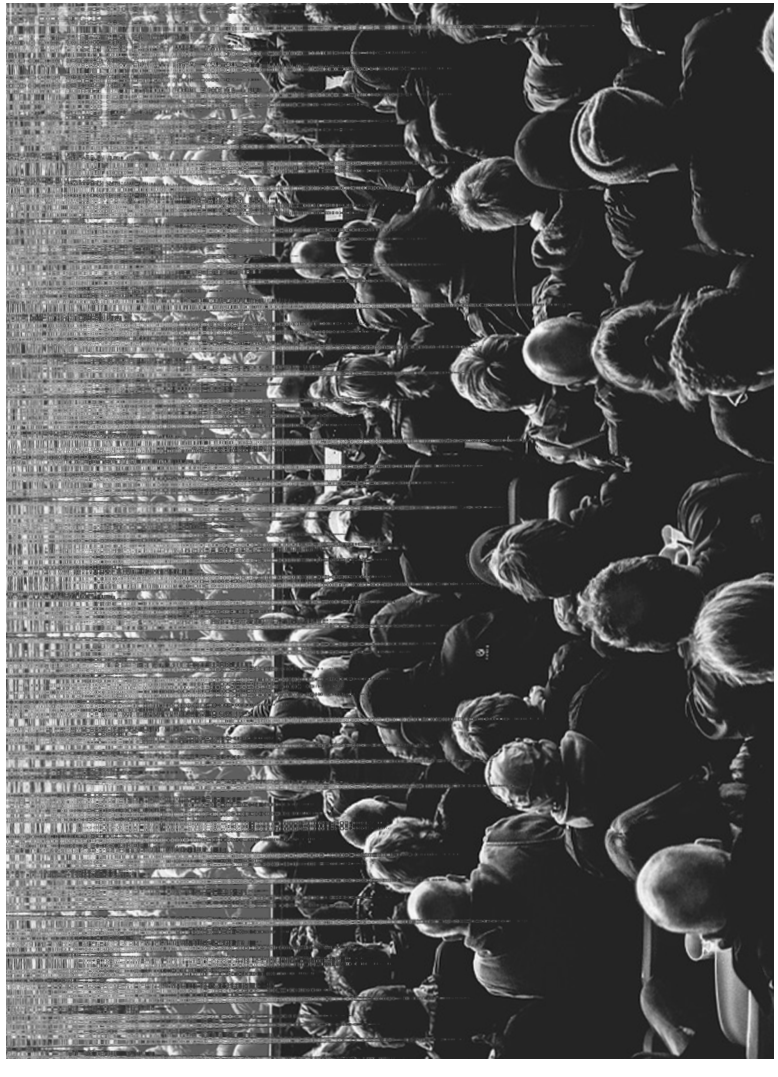
MARY DOES NOT LEARN ANYTHING NEW:

APPLYING KIM'S CRITIQUE OF MENTAL CAUSATION TO THE KNOWLEDGE ARGUMENT AND THE PROBLEM OF CONSCIOUSNESS



ADAM KHAYAT

ABSTRACT

Within the discourse surrounding mind-body interaction, mental causation is intimately associated with non-reductive physicalism. However, such a theory holds two opposing views: that all causal properties and relations can be explicated by physics and that special sciences have an explanatory role. Jaegwon Kim attempts to deconstruct this problematic contradiction by arguing that it is untenable for non-reductive physicalists to explain human behavior by appeal to mental properties. In combination, Kim's critique of mental causation and the phenomenal concept strategy serves as an effectual response to the anti-physicalist stance enclosed within the Knowledge Argument and the Zombie Thought Experiment.

The viability of mental causation in the discourse of mind-body interaction is an assumed tenet in psychology. This is also intimately associated with non-reductive physicalism, which holds that though everything can be explained via reduction to physics, there are multiple methods of describing physical reality. Therefore, various areas of the specials sciences—such as psychology, economics, and biology—are more abstract and have the capacity to satisfy certain descriptive and explanatory interests that fundamental physics cannot.[1] This approach is relevant to the philosophical discussions surrounding the Knowledge Argument and the significance of zombies. The Knowledge Argument, as presented by Frank Jackson, claims that conscious experience necessitates non-physical properties.[2] First put forth by David Chalmers, the Zombie Thought Experiment was constructed to elucidate issues concerning the relationship between consciousness and the physical world.[3]

Nevertheless, non-reductive physicalism seems to hold two opposing views: that all causal properties and relations can be explicated by physics and that special sciences have an explanatory role. Jaegwon Kim attempts to deconstruct this problematic contradiction by arguing that it is untenable for non-reductive physicalists to explain human behavior by appeal to mental properties. This paper is divided into three sections: Section I will discuss the principles of externalism, causal closure, and explanatory exclusion and how they pose problems for mental causation within a physicalist framework—they will also be applied to the Knowledge Argument; Section II will propose and critically appraise various rebuttals to the exclusion argument; Section III will attempt to apply Kim's reasoning to the Zombie Thought Experiment.

1 William Jaworski, *Philosophy of Mind: A Comprehensive Introduction* (Chichester: Wiley-Blackwell, 2011), 129.

2 Martina Nida-Rümelin, "Qualia: The Knowledge Argument," *The Stanford Encyclopedia of Philosophy*, last modified November 23, 2009, https://plato.stanford.edu/archives/sum2015/entries/qualia-knowledge/.

3 Robert Kirk, "Zombies," *The Standard Encyclopedia of Philosophy*, last modified March 16, 2015, https://plato.stanford.edu/archives/sum2015/entries/zombies/.

I.

Kim begins his argument with the following question: through what mechanism or process does a mental event manage to initiate, or insert itself into, a causal chain of physical events?[4] Such an inquiry stems from the ambiguity surrounding how mental states could directly influence neurophysical mechanisms. This intermingling of categorically different substances—as hypothesized by Descartes—would imply that the nonphysical mind must be able to affect the conditions of the physical mind; nevertheless, this interaction has not been adequately clarified.[5]

Conversely, one of the most discussed arguments against physicalism—the Knowledge Argument—is predicated upon the notion that complete physical knowledge of a conscious entity would not also encompass the experience of being that entity. Frank Jackson presents this with a thought experiment about a brilliant scientist named Mary:

1. Mary understands all the neurophysical information regarding human color vision before her release from a monochrome environment.
2. However, before her release, there is some information that she lacks concerning the subjective experience of color.
3. Therefore, not all information is physical information.[6]

The purported conclusion derived from this example is that there are certain truths regarding the subjective experience, or qualia, of seeing red that escapes the neurophysical one; thus, physicalism is incomplete.

As a thesis about semantic content, externalism serves as a significant challenge to the utilization of mental causation within a physicalist framework to explain behavior. Accordingly, the process of individuating mental states requires consideration of the physical environment and the linguistic standards of one's surrounding community.[7] Thus, the content of intentional states is extrinsic.[8] This is problematic in that causation is intuitively understood to involve intrinsic features. Consequently, the externalist ways of characterizing the content of mental states makes them unsuitable for causal involvement.[9] When contextualized within psychology,

4 Jaegwon Kim, *Philosophy of Mind* (Westview Press, 1996), 439.
5 Kim, *Philosophy of Mind*, 443.
6 Frank Jackson, "Epiphenomenal Qualia," in *Mind and Cognition: An Anthology*, ed. William G. Lycan and Jesse J. Prinz (MA: Blackwell Publishing, 2008), 659.
7 Hilary Putnam, "The Meaning of 'Meaning'," *Minnesota Studies in the Philosophy of Science* 7 (1975): 131-193.
8 Kim, *Philosophy of Mind*, 445.
9 Julie Yoo, "Mental Causation," *The Internet Encyclopedia of Philosophy*, last modified 2006, https://www.iep.utm.edu/mental-c/#SH3bii.

the contents of beliefs, desires, and other propositional attitudes have no causal relevance; only the syntax is of significance.[10] Kim illustrates this by presenting a classic example: Putnam's Twin-Earth Thought Experiment. Accordingly, an Earthling refers to H_2O upon utterance of "water" whereas a Twin-Earthling refers to XYZ—a superficially identical yet compositionally different substance—upon the same utterance. They behave in an identical fashion; however, when contemplating ideas about what they both call "water," they are thinking about distinct things. The culminating conclusion is that the meanings of words are not holistically psychological—the content of mental states does not completely depend on intrinsic properties.[11] Therefore, mental states—which depend on extrinsic properties—lack causal relevance. With regards to Jackson's thought experiment, it can be argued that the qualia from experiencing the various colors does not have any residual effect on Mary's behavior.

Another argument against mental causation within a physicalist framework revolves around the causal closure of the physical domain. Cartesian interactionism postulates that both mental and physical events can occur as links within the same causal chain. To the contrary, physicalism is committed to the proposition that the only causes are physical causes; everything can be exhaustively described and explained by physics.[12] Kim asserts that a physicalist must reject the mental realm as an ontological equal of the physical realm. Therefore, mental causation must be ruled out.[13] Nonetheless, non-reductive physicalists maintain that certain systems can have irreducible mental properties. Such a position is at odds with physicalism and is thus unsustainable.

The problem of explanatory exclusion emerges from the non-reductive physicalist view that mental causes are distinct from physical causes. Instantiations of mental properties are associated with particular physical properties. These physical properties can be seamlessly integrated within a causal chain that produces behavioral effects; however, this seems to make mental properties to be causally stagnant and thus "excluded" from causal explanation.[14] Moreover, if both the mental property and the physical property are said to be causal, a case of overdetermination results.[15] This seems to contravene the "maxim of explanatory simplification," which seeks to explain behavior

10 Kim, *Philosophy of Mind*, 452.
11 Putnam, "Meaning of 'Meaning'."
12 William Jaworski, "Mental Causation from the Top-Down," *Erkenntnis*, 65, no. 2 (2006): 68.
13 Kim, *Philosophy of Mind*, 453.
14 John Heil and David Robb, "Mental Causation," *The Stanford Encyclopedia of Philosophy*, last modified October 10, 2018, https://plato.stanford.edu/archives/win2018/entries/mental-causation/.
15 Kim, *Philosophy of Mind*, 455.

with the fewest posited premises.[16] Additionally, it is unclear how a certain mental state can catalyze a series of neurophysical mechanisms. This problem can be best articulated by the following set of jointly inconsistent claims:

1. Actions have mental causes.
2. Actions have physical causes.
3. Mental causes and physical causes are distinct.
4. An action does not have more than one cause.[17]

Claims 1 through 3 indicate that actions can have multiple causes, whereas Claim 4 suggests that they do not. Consequently, for a non-reductive physicalist, denunciation of one of the claims is needed to maintain argumentative coherence; however, this is difficult. Rejecting the first claim would be problematic for a non-reductive physicalist: to deny the existence of mental events or their causal influence seems to contradict the theory's basic premises. Rejecting the second claim would be at odds with non-reductivists' commitment to physicalism.[18] Rejecting the third claim would contradict non-reductivists' commitment to anti-reductivism.[19] Rejecting the fourth claim insinuates that events can be causally overdetermined. According to Kim, either mental events are realized by, or supervene on, physical events; nevertheless, in both cases, mental events require physical events to exist and therefore both cannot provide independent and fully satisfactory causes for actions.[20]

These components of Kim's argument are relevant to discussions surrounding the Knowledge Argument. By function of the principles of externalism, causal closure, and explanatory exclusion, the qualia associated with seeing color for the first time does not confer novel information regarding the neurophysical facts of human vision. The content of the derived qualia lacks causal significance.

Kim discusses various models that could explicate the role of mental events.[21] The epiphenomenalist model asserts that mental states are mere byproducts of neurophysical states and lack any causal role. The model of supervenient causation views mental states as a potential cause due to its supervenience on neurophysical states. The reductionist model—which Kim considers to be the most efficacious and simple—identifies mental states with neurophysical states, which function as the only stimulus for other physical states.

16 Jaegwon Kim, "Mechanism, Purpose, and Explanatory Exclusion," *Philosophical Perspectives* 3 (1989): 93.
17 Jaworski, "Mental Causation from the Top-Down," 170.
18 Jaworski, "Mental Causation from the Top-Down," 171.
19 Jaworski, "Mental Causation from the Top-Down," 129.
20 Jaworski, "Mental Causation from the Top-Down," 172.
21 Kim, *Philosophy of Mind*, 455.

II.

Though the contentions proposed by Kim are challenging for non-reductive physicalism, there are various responses.

One potential reply to the exclusion argument falls within the designation "autonomy solutions." As previously discussed, Kim argues that non-reductive physicalism is guilty of pitting higher-level mental properties against their corresponding lower-level neurophysical properties in determining causation.[22] However, according to some, this description is reductionist and deceptive; rather, psychological justifications—and others in the special sciences—are independent of physical explanations in that they refer to their own collection of rules and abstract away from the details of physical explanations.[23] Subsequently, exclusion of mental causation can be prevented within a physicalist framework; both descriptions can coexist.

This approach is best illustrated by the dual explanandum strategy. Accordingly, purely physical explanations of behavior are not capable of satisfying certain explanatory interests: why Syria is engulfed in conflict cannot be explained via the interaction of molecules and atoms. In his argumentation, Kim produces a paradox of psychological explanation: he claims that psychological explanations—which refer to mental states—lack objective status and are excluded by neurophysical explanations, which, in turn, are incapable of fulfilling explanatory interests that fall within the domain of psychological explanations.[24] An unappealing ultimatum results: either neurophysical statements—which can elucidate objective relations but cannot answer certain special questions—or psychological statements—which can answer certain special questions but cannot describe objective relations—are accepted. In contrast, the dual explanandum strategy revolves around the notion that causation cannot be extricated from the explanatory schemes in which it functions.[25] It argues that the causal relations that emerge from psychological explanations and neurophysical explanations serve different purposes; they form mutually exclusive and autonomous causal lines that are relevant to different properties of the end effect.[26] Neurophysical explanations can describe the interaction of atoms, whereas psychological explanations can clarify "the successful or unsuccessful interaction of organisms with their natural, historical and cultural environment."[27]

22 Heil and Robb, "Mental Causation."
23 Frank Jackson, "Mental Causation," *Mind* 105, no. 419 (1996): 386.
24 Karsten R. Stueber, "Mental Causation and the Paradoxes of Explanation," *Philosophical Studies* 122, no. 3 (2005): 256.
25 Yoo, "Mental Causation."
26 Heil and Robb, "Mental Causation."
27 Stueber, "Mental Causation and the Paradoxes of Explanation," 256.

This separation can be described by differentiating "triggering" and "structuring" causes. The former refers to the mechanism by which a particular effect is induced and lies within the purview of neurophysical explanations; the latter refers to the motive for why a particular effect is induced and lies within the purview of psychological explanations.[28] For example, the thermostat activates the furnace due to a low external temperature (triggering), but the organization of the circuitry forms the pre-conditions (structuring) that enable the low temperature to exert its effect.[29] With respect to the human mind, the external conditions that galvanize bodily behavior are mediated by the agent's learning history.[30]

Nonetheless, it has been argued that this strategy violates the causal closure of the physical domain. By claiming that certain aspects of the final effect can be attributed to causes that are irreducibly mental, the proposition of physicalism is infringed.[31] If this is not the case, then the problem of exclusion persists.[32] This is an unsustainable position for a non-reductive physicalist to hold.

Another reply to the exclusion argument is classified as the "inheritance solution." This is derived from a problem associated with Kim's critique: a property needs to be causally efficacious in the process of production for it to be considered as causally relevant to the production of certain effects.[33] Accordingly, what deems a property to be causally efficacious is that its instantiation leads to the manifestation of the effect. However, it has been argued that a distinction can be made between that which is causally relevant and that which is causally efficacious. From this, it follows that a psychological explanation is inefficacious but relevant because "its realization programs for the realization of a lower-order efficacious property and, in the circumstances, for the occurrence of the event in question."[34] In other words, it acquires this causal relevance due to its close interaction with its neurophysical realizer.[35] Psychological explanations and neurophysical explanations are not in competition for causation but are rather in cooperation; this circumvents the problem of causal overdetermination.[36]

Though this solution seems to be effective, counterarguments can be proposed. One is that causal inheritance is simply a form of

28 Fred Dretske, "Reasons and Causes," *Philosophical Perspectives* 3 (1989): 10.
29 Dretske, "Reasons and Causes," 11.
30 Yoo, "Mental Causation."
31 Kim, "Mechanism, Purpose," 101.
32 Yoo, "Mental Causation."
33 Frank Jackson and Philip Pettit, "Program Explanation: A General Perspective," *Analysis* 50, no. 2 (1990): 111.
34 Jackson and Pettit, "A General Perspective," 115.
35 Frank Jackson and Philip Pettit, "Functionalism and Broad Content," *Mind* XCVII, no. 387 (1988): 399.
36 Heil and Robb, "Mental Causation."

supervenience causation.[37] Accordingly, the mental property that is realized by its neurophysical counterpart is neither necessary nor informative; reduction is the only solution. Another identifies causal inheritance as a mere appeal to epiphenomenalism.[38] Since causal inheritance credits causal relevance to mental properties in virtue of their physical realizers, they "mental properties" have no inherent and independent causal power. Thus, they are superfluous.

The various replies to the exclusion argument have not been effective. They seem to indirectly appeal to mental causes as being ontologically equal to neurophysical causes in that they affect some aspect of the end result. Such an implication contradicts the basic thesis of physicalism. Kim's argument is successful in indicating that the non-reductive physicalist position is unmaintainable by analyzing the contradictory nature of its fundamental premises. It is also noteworthy in that its assertions are relevant to multiple philosophical contexts. With respect to the problem of consciousness in a physical world, implementation of Kim's approach begets interesting conclusions.

III.

An attempt at refuting the viability of physicalism, the Zombie Thought Experiment is founded upon a simple hypothetical scenario: there exists a system that is physically identical to a conscious entity but lacks that consciousness completely.[39] Such an approach presents a significant challenge to physicalism in that it obeys the casual closure of the physical domain, yet maintains that a fully physical account is insufficient: it does not describe how it is "to be like" something. From this, Chalmers differentiates between the "easy" and "hard" problems of consciousness.[40] Accordingly, the former revolves around the neurophysical processes that underlie discerning stimuli, reporting information, or assessing internal states; such activities embody puzzles that can be deciphered via empirical investigation. However, the latter involves the difficult prospect of reconciling the existence of qualia with a neurophysical description of the mind; an effective solution would require an explanation of the relationship between neurophysical processes and consciousness on the basis of natural principles. Chalmers introduces an epistemic argument to ground his contentions:

37 Jaegwon Kim, "Blocking Causal Drainage and Other Maintenance Choices with Mental Causation," *Philosophy and Phenomenological Research* 67, no. 1 (2003): 171.
38 Ivar Hannikainen, "Questioning the Causal Inheritance Principle," *Theoria: An International Journal for Theory, History and Foundations of Science*, SEGUNDA EPOCA 25, no. 3(69) (2010): 275.
39 David Chalmers, *The Character of Consciousness* (Oxford: Oxford University Press, 2010), 108.
40 Chalmers, *Character of Consciousness*, 105.

1. There is an epistemic gap between neurophysical and phenomenal states.
2. If there is an epistemic gap between neurophysical and phenomenal states, then there is an ontological gap, and materialism is false.
3. Thus, materialism is false.[41]

He then delineates three argumentative avenues by which materialists can oppose the epistemic argument: Type-A, Type-B, and Type-C materialism.

Type-A materialism flatly denies Claim 1: there is no epistemic gap between neurophysical and phenomenal states. This approach denies the existence of consciousness and phenomenal states; descriptions of neurophysical processes can exhaustively explain human behavior. Chalmers suggests that such a stance is extremely counterintuitive and lacks a strong argument.[42]

Type-B materialism accepts that there exists an epistemic gap but rejects Claim 2: there is no ontological gap. Consequently, phenomenal states can be identified with neurophysical states.[43] An example of this would be the identification of water with H_2O. Nevertheless, Chalmers claims that this approach is untenable; the epistemic gap with consciousness seems to be distinct from epistemic gaps in other domains. In other words, the identification between consciousness and neurophysical states is "epistemically primitive." The identity is not deducible from the complete physical state.[44]

Type-C materialism also accepts that there exists a deep epistemic gap between neurophysical and phenomenal states but claims that such a gap is closeable with further empirical investigation. Therefore, phenomenal states are deducible in principle from physical states, but these inferences are unavailable now.[45] Chalmers contests the plausibility of this argument via a categorical approach. By not designating consciousness as a functional concept and by classifying physical descriptions of the world as structural-dynamic descriptions, Chalmers asserts that consciousness cannot be implied by a neurophysical description. Thus, either Type-A or Type-B materialism can be accepted; there is no distinct space for Type-C.[46]

Kim's argument is relevant to this discussion in that his approach seems to embody a hybrid between Type-B and Type-C materialism and—via the principle of explanatory exclusion—rids consciousness

41 Chalmers, *Character of Consciousness*, 112.
42 Jaworski, *Philosophy of Mind*, 215.
43 Chalmers, *Character of Consciousness*, 117.
44 Chalmers, *Character of Consciousness*, 118.
45 Chalmers, *Character of Consciousness*, 126.
46 Chalmers, *Character of Consciousness*, 130.

or qualia from any causal role in the physical realm. When combined with the phenomenal concept strategy, Kim's reasoning can thus serve as a successful response to the Zombie Thought Experiment because it leads Chalmers to accept epiphenomenalism as the only viable solution. Subsequently, epiphenomenalism fails to integrate with a naturalistic worldview and to respond to the problem of psychophysical emergence.[47] Therefore, Chalmers' argument seems unsustainable.

As aforementioned, Kim advocates for a reductionist model, which seems to be the most efficacious. Instantiations of consciousness, or qualia, can be reduced to corresponding neurophysical mechanisms. This adheres to the basic tenets of Type-B and Type-C materialism; with further empirical investigation, the ambiguities surrounding phenomenal states will be explicated by neurophysical processes. Furthermore, as previously indicated, it is unclear how consciousness, or qualia, would be able to exert an effect within the physical realm. The Zombie Thought Experiment seems to reinforce this by inverting the archetypical issue of mental causation; if zombies are physical duplicates that behave in an identical manner yet lack qualia, then qualia have no role in affecting or determining behavior.[48] Epiphenomenalism—or what Chalmers designates as Type-E dualism—is the best available option.[49] Consequently, there exists only an epistemic gap between neurophysical and phenomenal states.

The phenomenal concept strategy serves as a valid challenge to Claim 2: the inclusion of an epistemic gap does not necessarily imply the existence of an ontological one. Chalmers forcefully claims that physicalism denies "the manifest" and the "further truth that we are conscious." It can be argued, however, that phenomenal states simply assume the presence of ontologically separate and non-physical entities.[50] According to this approach, physicalists can accept the conceivability of zombies while insisting that consciousness, or qualia, is a conceptually isolated phenomenal concept which is intrinsically related to the neurophysical.[51] For example, there exists legitimate skepticism regarding the obviousness of qualia as ontologically separate from the standpoint of psychological language. If terms such as "pain" referred to private, subjective experiences, it would be expected that the derivation of the appropriate use of the term would only occur via introspection.[52] Nevertheless, conceptual analysis of psychological language reveals that the usage of the word "pain" is learned by

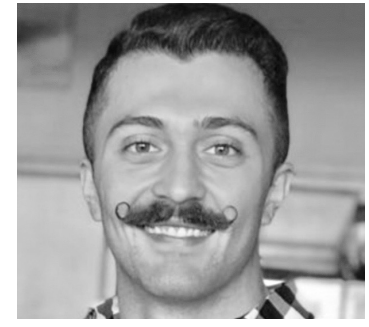
47 Jaworski, *Philosophy of Mind*, 229.
48 Kirk, "Zombies."
49 Chalmers, *Character of Consciousness*, 144.
50 Peter Carruthers and B. Veillet, "The Phenomenal Concept Strategy," *Journal of Consciousness Studies* 14, no. 9-10 (January 2007): 212-36.
51 Kirk, "Zombies."
52 Jaworski, *Philosophy of Mind*, 217.

associating linguistic behavior with non-linguistic behavior; it is not a private process.[53] Moreover, by appeal to ontological naturalism, consciousness, or qualia, seems to not exist; if the behavior of the zombie can be exhaustively described by neurophysical mechanisms and is indistinguishable from its duplicate, then qualia may not exist. Additionally, it is unclear how such phenomenal states could emerge from neurophysical states. If there exists a gap between neurophysical and phenomenal descriptions, how would it be possible for neurophysical processes to give rise to consciousness, or qualia? There is no appropriate response from epiphenomenalism.[54]

Furthermore, the phenomenal concept strategy would offer an appropriate response to Claim 2 of the Knowledge Argument. As opposed to learning a new fact—regarding color—that operates outside the neurophysical description, Mary simply understands an "old fact in a new way"—i.e. she has acquired a phenomenal concept of a neurophysical mechanism. Ultimately, this phenomenal concept can be explicated in neurophysical terms.[55] Consequently, phenomenal concepts can be reduced to physical properties of experiences.

In combination, Kim's critique of mental causation and the phenomenal concept strategy serves as an effectual response to the anti-physicalist stance enclosed within the Knowledge Argument and the Zombie Thought Experiment. By demonstrating that subjective experience, or qualia, is causally inert and is not ontologically independent, this approach pushes advocates of anti-physicalism to accept epiphenomenalism as the only viable alternative. In turn, epiphenomenalism suffers from an inability to integrate with a naturalistic worldview and to respond to the problem of psychophysical emergence. Thus, the contention that the supposed existence of consciousness is sufficient reason for the failure of physicalism is not successful.

53 David W. Schaal, "Naming Our Concerns about Neuroscience: A Review of Bennett and Hackers Philosophical Foundations of Neuroscience," *Journal of the Experimental Analysis of Behavior* 84, no. 3 (2005): 683-92.
54 Jaworski, *Philosophy of Mind*, 239.
55 Nida-Rümelin, "Qualia: The Knowledge Argument."



ABOUT THE AUTHOR:
Adam Khayat, a senior at the University of Louisville in Kentucky, is majoring in biology and minoring in philosophy. His primary philosophical interests are postcolonialism, philosophy of mind, bioethics, existentialism, and Islamic philosophy. In the fall, he will be attending medical school.