VERTAISARVIOITU
KOLLEGIALT GRANSKAD
PEER-REVIEWED
www.tsv.fi/tunnus

19
On the irrelevance of freedom to the causal relevance of will

Renne Pesonen

Introduction

Assume that there is no free will. Does it follow that everyone can do as they please, since no one can be held accountable for their actions? Many find the affirmative answer intuitive. Some of them embrace moral nihilism, but I believe most philosophers who consider the affirmative answer correct argue that we must assume the existence of free will because we indeed are morally accountable for our actions.

There is also an analogous—and maybe deeper—question concerning the ontological status of actions themselves: If there is no free will, does it follow that no intentional action is possible, since our behavior is predetermined, perhaps by the laws of nature, regardless of our beliefs, desires, and reasons? Or, to rephrase: if intentional action is indeed possible, does it follow that there is free will? In this essay, I examine two views that accept this inference but arrive at opposing stances on the question of free will.

The first view denies the existence of free will based on reductive physicalism. Since everything is physical and every event is determined by the laws of physics, there can be no true causal principles or powers that transcend the causal closure of the physical. Any putative higher-level causation would violate the fundamental

Pesonen, R. (2025). On the irrelevance of freedom to the causal relevance of will. In A. Laitinen, M. Keinänen, J. Reinikainen & A. Honkasalo (eds.). *Language, Truth, and Reality*. Tampere University Press. (pp. 237–247). https://doi.org/10.61201/h5fbdk64

laws of nature. In particular, free will is impossible because mental causation cannot exist. I call this stance *metaphysicalism*. Its adherents may claim that their view is based on natural science, but it is a theory of metaphysics, rather than physics, that licenses their inferences. The gist of their view is the causal closure of the physical and the reducibility of every thing, property, and event to physics. However, it does not ultimately matter which theory of physics is correct and what the actual laws are—for example, whether they are deterministic or not.

I call the second view *intentional realism*, which refers to realism about the ontology of intentional ascriptions. Intentional realists rely on science, but it is the behavioral, cognitive, and social sciences rather than physics. According to them, intentional ascriptions may not constitute proper scientific laws, but they are nevertheless indispensable for explaining and predicting human behavior. Moreover, if our best theories of psychology, economics, sociology, and so on, require that people make and enact choices among alternative possibilities, we are licensed to presume the existence of free will (or its psychological cognates). Thus, intentional realists are scientific realists who may or may not subscribe to the thesis of causal closure of the physical. What they insist on is that mental or intentional states are not superfluous but have explanatory and predictive relevance.

Below, I formulate a defense of intentional realism based on standard arguments for anti-reductive physicalism and the autonomy of the special sciences. While the argument salvages the causal efficacy of will from the ultimately absurd and militantly anti-scientific attack of metaphysicalism, it implies nothing about the question of freedom in terms of the metaphysical possibility of genuine alternatives. I argue that the metaphysics of freedom should be disentangled from the scientific questions concerning the role of will (or related mental states) in intentional explanations. Whether or not the will is free in any metaphysically or morally relevant sense is ultimately irrelevant to explanations and predictions that invoke mental causation.

Reasons for anti-reductionism

Once we look at the actual theories and practices of science, it immediately becomes evident that metaphysicalism is an anti-scientific ideology. Even if everything is material and governed by the laws of nature, the theories and methods of physics cover only a small portion of legitimate scientific questions. If you want to know whether rising interest rates mitigate inflation or whether schizophrenia is heritable, you need other methods and theories to answer these questions. Denying the reality of these phenomena, or the legitimacy of any means beyond physics to investigate them, clearly represents a doctrine of militant anti-science. The next section discusses psychological explanation and metaphysicalism. However, let's first take a glance at scientific reductionism, which is a less militant and more general program in comparison to metaphysicalism.

Back in the day, positivists argued that it should be possible to reduce the entities and laws of "higher-level" sciences, such as biology, to lower and more fundamental levels, such as chemistry, and ultimately all the way down to fundamental physics (e.g., Oppenheim & Putnam, 1958, Nagel, 1961). This branch of reductionism did not necessarily deny the existence of higher-level entities or laws. Rather, it aimed to show that, in principle, they could be derived from, or explained by, the laws of physics. However, in recent decades, few philosophers of science have subscribed to this program. Most of us are *non-reductive* physicalists now.

Non-reductive physicalists believe in metaphysical materialism and the causal closure of the physical, but they also maintain that there are higher-level (such as emergent or supervenient) properties and causal regularities that cannot be defined or derived from theories of physics. For example, a 50-euro bill is obviously a physical thing, but its value or status as legal currency does not stem from its physical properties. Every financial transaction is a physical event, but there are unlimited ways to physically implement these transactions—whether through gold coins, fiat bills, exchanges of information in computer networks, or whatever the future may bring. Moreover, aggregates of those transactions form higher-level social and economic patterns that sometimes can be predicted and manipulated. The fact that prices tend to go up as demand increases is not something that can be derived from fundamental physics. Furthermore, we do not need physics to investigate whether this pricing trend can be reversed by increasing production. Discoveries concerning the laws of fundamental physics will almost certainly have no impact on theories in fields such as macroeconomics, population genetics, or psychology.

In a nutshell, the above is the standard argument put forward, for example, by Fodor (1974, 1997), in support of non-reductive physicalism and the autonomy of the "special" sciences. While the argument is traditionally formulated in terms of theories and laws, this is not necessary. Instead, it is common to conceive of scientific explanations in terms of variables and dependencies between their values (Woodward, 2000): If you change the value of variable X, the value of Y tends to change. If this relationship is an established invariance, you can predict and potentially explain the values of Y based on changes in X. For example, if you increase the intake of vitamin C in a malnourished population (or the production of a good in a market), the incidence of scurvy (or the price of the good) decreases. Using such regularities for prediction and scientific explanation does not require them to be laws in any strict sense. Non-reductive physicalists further argue that these kinds of causal generalizations cannot be reduced to lower-level sciences if the higher-level mechanisms and properties that make them work can be physically realized in multiple ways.

Functional explanations in psychology

Since the advent of functionalism in the philosophy of mind (Fodor, 1968; Putnam, 1967; Block & Fodor, 1972), it has been commonplace to hold that psychological states are prime examples of multiply realizable phenomena. Our beliefs and desires are somehow realized in our brains, but they also have characteristic consequences for both mental and overt behavior, which can be identified and investigated without knowledge of their physical realization. According to functionalists, it is these characteristic consequences, identified at the intentional rather than the physical level of causation, that make them instances of psychological states such as beliefs, desires, and so on.

Following Dennett's (1971) classic analysis, consider, for example, a computer running a chess program. If the program's behavior is rational enough, we can explain its moves simply by referring to the strategies and rationales of the game. The same qualitative behavior can be implemented by infinitely many algorithms, and we do not need to know the exact algorithm to explain or predict the moves. Furthermore, the same algorithm can be executed on physically vastly different machines, and a complete description of the machine would not usually help in predicting the program's behavior, simply due to its excessive complexity. Likewise, with humans, we do not need to know much about the exact mental machinery, and even less about the brain, to predict and explain people's everyday rational behavior. Intentional explanation is justified purely by explanatory and pragmatic necessity. Moreover, there is no extra mystery in the relationship between the mind and the brain compared to that between a program and the machine.

With these teachings in mind, I discuss an example borrowed from Raatikainen (2010): Suppose John desires a bottle of beer and believes that there is some in the fridge. Without any knowledge of his brain, we can safely bet that, soon enough, he will head to the fridge. However, before John gets a chance to get there, we tell him that we already drunk all the beer and the fridge is empty. Hence, John's belief changes from "there is some beer in the fridge" to "there is no beer in the fridge." This intervention changes his behavior from "go to the fridge" to "go to the store to get more beer." While changes in John's beliefs and behavior certainly implicate changes in his brain states, we can explain the change in his behavior solely in terms of changes in his beliefs. We have absolutely no information on exactly how his brain was affected, and we do not need that information in order to make inferences about his behavior.

Raatikainen (2010) argued that since psychological states are multiply realizable, we could, in principle, manipulate John's brain states without manipulating his beliefs concerning the beers and the fridge. Such an intervention would not affect John's decisions and behavior, while an intervention on his relevant beliefs would. What follows is that, surprisingly, it is the changes in intentional states that explain change in his behavior, not changes in the brain!

I think that it is largely correct. However, Raatikainen uses the thesis of the multiple realizability of the mental in a slightly non-standard and potentially problematic way. The standard view holds that mental states and processes could, in principle, have vastly different realizations *beyond* the (typical) human brain. This is because, if mental states are identified functionally—based on their causal properties concerning perception, action, and other mental states—similar causal networks could, in principle, be realized in vastly different brains or even without brain tissue at all, for example, in silicon-based life forms or robots. This is not the same as claiming that changes in an individual brain could occur without changes in behavior. Therefore, one could accept the antecedent of the argument without accepting the consequent, because not just any intervention on the brain should count as relevant, regardless of whether mental states are multiply realizable.

We could, for example, manipulate the firing rates of some random neurons at the periphery of John's visual cortex, and no one expects this would change his behavior. Therefore, we need a specification of which interventions count as relevant. Raatikainen (2010, 359) argues that, in the given example, the only route through which we can change John's behavior is by altering his beliefs. I think this is almost correct. However, if we allow direct interventions in John's brain, it should be possible to stimulate his motor areas to make him head for the grocery store instead of the fridge without altering his beliefs or desires. Hence, behavior may not always align with beliefs and desires, and it could be argued that the brain is causally explanatory for John's behavior after all.

However, functionalists identify psychological states with *patterns* of behaviors and inferences. If we only hijack John's motor cortex, he presumably still believes there is beer in the fridge and remains disposed to make inferences based on that belief. He may still want to go to the fridge, but his behavior is now under external control, and he is acting in a way that he cannot rationally explain. Our intervention has interfered with the normal functioning of John's brain and mind. When this happens, we can no longer explain his behavior in terms of intentional states because their normal causal functions no longer exist. Such an intervention would not count as an argument against the causal efficacy of mental states, as the relevant causal mechanisms are severed.

What if we hijack a part of John's brain so that we shift the entire *pattern* of behavior and inferences stemming from his beliefs about the beer and the fridge? That sort of intervention surely counts as relevant. However, according to functionalists, such an intervention would amount to altering John's beliefs, because beliefs are identified precisely by such patterns. That intervention would not rob John of his capacity to act freely or at least rationally. We routinely create such interventions by simply telling people that there is no more beer in the fridge and so on. This slight alteration of Raatikainen's (2010) argument, based on the functionalist theory of mental states rather than on mere multiple realizability, retains its original conclusion.

However, not everyone believes in functionalism, and perhaps we could dispense with belief/desire explanations altogether. Maybe we could scan John's brain and

use theoretical calculations (based on some future neuroscience) to make even better predictions about his behavior. Well, perhaps we could, yet we don't. We know how intentional explanations work, but we have absolutely no idea how to predict complex behavior from brain activity, let alone from fundamental physics. Commonsense belief/desire ascriptions are surely too blunt an instrument for serious scientific psychology, but we routinely rely on them for quotidian and social scientific explanations. Swaths of cognitive psychology, behavioral economics, and related fields operate on the intentional level of explanation, even when they attack our folk psychological platitudes. I am not here defending functionalism as such but scientific realism in the realm of intentional explanation.

I close this section by outlining what the metaphysicalist alternative would be. What would John do if he believed that there was beer in the fridge that he desires? That depends on the laws of nature and the elementary particles that comprise his body. The particles do not care about John's desires, so we would need to know the complete physical description of his body and perform extremely complex calculations to predict what is going to happen. I am not sure if such predictions are possible even in principle, but that is the only option available for metaphysicalists. Furthermore, once we tell John that the fridge is empty, what would follow? The imagined situation leaves John's physical state and the impact of the intervention completely unspecified. Therefore, according to metaphysicalists, absolutely no predictions follow, and nothing in the given description could explain John's behavior.

Consistent metaphysicalists should not even care about arguing that it is the brain that drives behavior. Brain processes are surely physical, but for metaphysicalists, there are no higher-level properties or causal regularities to be identified beyond fundamental physics. In the end, this tenet does not render only special sciences impossible but science itself, including physics. If metaphysicalism is true, no one would conduct experiments for epistemological reasons or accept hypotheses for rational reasons. In fact, metaphysicalists themselves would not hold their beliefs because it is the rational thing to do, but simply because things turned out that way.

So much for the metaphysicalism. This section conveys three main points: (1) "The brain made me do it" may, in a sense, always be a correct answer when you need to explain your actions. However, it almost never is the only correct option or the most informative one. (2) The debates between intentional versus other scientific explanations of behavior (such as brain centered explanations) are not the same as the debate between metaphysicalism and intentional realism. This is further discussed below. (3) Nothing said thus far bears on the question of freedom in terms of metaphysically possible alternatives. Intentional states can be causally efficacious or relevant even if there is no freedom in that sense. With these conclusions in mind, I next discuss potential misconceptions about determinism, consciousness, and external influences in debates concerning the freedom of will.

Some implications of the argument

Some enemies of free will mount their attack from neuroscience or biology. It is your brain or genes calling the shots, so your thoughts and will are mere illusions, or at least their presumed significance for your behavior. These arguments may look much like metaphysicalism, but instead of radical materialism, they rely on behavioral sciences to argue that our actions are determined by external or non-conscious factors beyond our control. While I cannot discuss these debates in detail here, I will briefly address some confusions they harbor concerning these factors in relation to intentional explanations. I believe that at least some misunderstandings can be straightened out simply by disentangling questions concerning freedom from an entirely different question about the role of will or mental states in psychological explanations.

The determinism/non-determinism dimension

Incompatibilists believe that free will requires the universe to be indeterministic, for if every event is determined by the laws of nature, there is simply no room for the will to operate. This topic veers back into metaphysical debates, on which I have nothing more to say. However, if we accept the autonomy of intentional explanations, we can have indeterminism without metaphysics, simply because intentional ascriptions are not fundamental laws but probabilistic generalizations.

For example, if we tell John that there is no beer in the fridge, nothing in the intentional description determines what he will do next. Maybe he still goes to the fridge to check. Maybe he skips the trip to the grocery because he doesn't bother. Typically, we use intentional ascriptions for explanation rather than prediction, and often they serve merely as *post hoc* rationalizations (see Cushman, 2019). However, this does not mean that they cannot factor into legitimate scientific explanations.

The aims of science are diverse, and the aims of explanation and understanding are not always aligned with the aim of prediction (Potochnik, 2015). That is why many explanatory models in science abstract and idealize. What intentional ascriptions capture are not laws; they provide only an approximate model for explaining human behavior, which is highly patterned but still not deterministic at the intentional level of description. Reasons, instincts, norms, and so on may be in conflict and they rather motivate than determine decisions. Perhaps what we experience as freedom of action is simply the result of a host of variables that render any particular action practically unpredictable. At any rate, indeterminism in the domain of psychology does not imply any metaphysical commitments, but neither does it imply randomness or a lack of control.

The internal/external dimension

Perhaps it is our genes or the environment that truly determines our actions. Indeed, instincts, habits, and social norms may often explain our behavior better than our conscious volitions (Cushman, 2019), and factors such as mental illness or coercion

can also rob us of the possibility to act freely. An unquestionably important fact is that many factors external to our conscious intentions guide our behavior, and we are not always as free as we believe or want to be.

However, these considerations do not imply anything fundamental about human freedom, except that the extent to which our choices are free is partly an empirical and partly a moral consideration. Many of the factors mentioned simply modulate our intentions, and they are comparable to beliefs and desires in that they affect our behavior without strictly determining it on each occasion. At gunpoint, you may still be free to make the rational decision to give up your wallet before your life. Your chess moves or your next steps when the fridge proves empty may be strongly habitual, but they are not strictly determined by your genes or your past any more than each move of a chess machine is hard-coded into its program.

I believe the future of behavioral sciences will push the boundaries of what we can predict and explain, and this will have consequences for people's intuitions about which behaviors are free and which are determined beyond our control. The point I want to make here is that the debate about the relative importance of external versus internal factors in the explanation of behavior is not the same as the debate between metaphysicalists and intentional realists. Metaphysicalists cannot even differentiate between internal and external causes because, for them, there are no identifiable higher-level determinants of behavior. For intentional realists, the question about the relative importance of external and internal factors is not a fundamental or moral question about human freedom but an empirical one concerning the causal variables involved in intentional explanations.

The consciousness/non-consciousness dimension

Finally, some regard consciousness as crucial for free will (e.g., Hodgson, 2012). I argue that it actually isn't, at least insofar as free will pertains to decision-making and cognitive control.

The most famous example of the dominance of unconscious over conscious brain processes is the Libet experiment (Libet et al., 1983). It demonstrated that a brain signal can be reliably detected before subjects experience a conscious volition to move their arm. Thus, in this simple task, the brain makes the decision first, and apparently the conscious intention follows. However, this is not very alarming.

Behavioral and decision scientists have held for decades that human decision-making is largely intuitive and often eludes conscious control. However, this does not mean we lack control or freedom over our decisions. For example, an influential dual-process model by Kahneman (2003) posits that unconscious processes make automatic decisions that we rely on during routine activities. The function of conscious processes is to monitor these decisions and ongoing behavior, intervening when there is a reason to do so. Therefore, simple and trivial decisions that do not require planning or control, such as those investigated by Libet et al., are expected to stem from non-conscious processes. The conscious mind simply monitors what is happening and exercises regulatory control when necessary. Importantly, dual-

process and related theories do not conceptualize the difference between non-conscious and conscious processes in terms of brain versus mental processes. Non-conscious processes may be automatic but still intentional. This is particularly clear when automatization results from habituation during social or skill learning.

But can an entirely unconscious mechanical system, such as artificial intelligence, make decisions? I don't see why not. The entire field of reinforcement learning investigates decision-making and learning in artificial agents, with many methods inspired by psychological and biological learning theories (Sutton & Barto, 2018). As far as I can tell, it should be possible to model human decision-making as accurately as we wish. It might be natural to view these non-conscious systems as mere complex automata, incapable of doing anything beyond what their programs dictate. However, if the human mind surpasses the capabilities of mere machines, consciousness is probably not among the reasons. The notion of consciousness used by dual-process and related decision theorists refers to access consciousness (see Block, 1995) rather than subjective experience. Access consciousness is a functional concept involved in metacognition. As such, it should be amenable to causal description like any other mental function.

For example, Hodgson (2012) believes that consciousness does not conform to any rules or laws, but it may contribute to free decision-making, for example, by resolving inconclusive reasons. If this means a capacity to form an arbitrary choice, it seems to me that the same function can be implemented simply by drawing decisions randomly from some appropriate distribution.

In conclusion, the fact that some action-guiding processes are not conscious does not mean that they are mere mindless brain processes. Moreover, consciousness is not essential for the ability to form and exercise one's will, insofar as it refers to the ability to rationally choose one's goals and actions.

Conclusion

I have argued that the problem of free will involves two separate questions. On the one hand, it involves the mind–body problem, which, in this context, boils down to the question of the causal relevance of mental states. On the other hand, there is the question of whether the will can be truly free. There are scientifically respectable answers to the first problem, and metaphysicalism is not among them. To me, the second question appears to be primarily moral or metaphysical. I do not wish to imply that the question is therefore meaningless, but rather that it should be disentangled from the question of the causal efficacy of the will. The remaining scientific question is not whether the will is free, but whether it plays a role in scientifically respectable explanations of behavior.

But does it? You rarely encounter the term "will" in the literature of cognitive or behavioral sciences. For the purposes of this essay, I have considered it to be a folk-psychological abstraction that captures aspects of executive function, such

as decision-making, goal selection, and cognitive control. As with other folk-psychological notions, such as beliefs and desires, it serves only as an approximate but heuristically useful description of the determinants of intentional behavior. If "will" cannot be given any psychologically meaningful functional interpretation, it is futile to debate whether it is free or real, as it would lack any explanatory relevance anyway.

I suspect that arguments from genes and the brain against the existence of free will appeal to some because they reveal the human mind to be a physical system, after all. However, functionalists do not deny this. Their point is simply that human behavior can also be described in terms of intentional agency, where causal variables are identified based on their causal properties in cognition and behavior rather than their intrinsic physical properties. It is all about levels of description that justifies the use of intentional explanations. Dennett's (1971) argument for intentional interpretations can be criticized on the grounds that it justifies anthropomorphizing machines when it is convenient to think of them as intentional agents. But surely we cannot be guilty of anthropomorphizing humans!

Hence, it is futile to argue against intentional realism on the basis that science shows the mind to be a complex causal mechanism. At least functionalists already believe this. For them, science is there to uncover the nature and exact mechanisms of mental functions, often correcting our preconceived folk-psychological ideas about them. As research progresses, it may very well chip away at our intuitive belief in human freedom. However, no harm is necessarily done. The facts uncovered thus far do not justify metaphysical nihilism or biological determinism; rather, they only place restraints on excessively libertarian conceptions of human freedom and reason.

References

Block, Ned (1995): 'On a confusion about a function of consciousness', *Behavioral and Brain Sciences,* 18(2): 227–247. URL = https://doi.org/10.1017/s0140525x00038188.

Block, Ned & Fodor, Jerry (1972): 'What Psychological States are Not', *The Philosophical Review,* 81(2): 159–181. URL = https://doi.org/10.2307/2183991.

Cushman, Fiery (2019): 'Rationalization is rational', *Behavioral and Brain Sciences,* 43(28). URL = https://doi.org/10.1017/S0140525X19001730

Dennett, Daniel (1971): 'Intentional Systems', *The Journal of Philosophy,* 68(4): 87–106. URL = https://doi.org/10.2307/2025382.

Fodor, Jerry (1968): *Psychological Explanation: An Introduction to the Philosophy of Psychology,* New York: Random House.

Fodor, Jerry (1974): 'Special sciences (Or: The Disunity of Science as a Working Hypothesis)', *Synthese,* 28(2): 97–115. URL = https://doi.org/10.1007/bf00485230.

Fodor, Jerry (1997): 'Special Sciences: Still Autonomous After All These Years', *Noûs,* 31(S11), 149–163. URL = https://doi.org/10.1111/0029-4624.31.s11.7.

Hodgson, David (2012): *Rationality + Consciousness = Free Will,* Oxford: Oxford University Press.

Kahneman, Daniel (2003): 'A Perspective on Judgment and Choice: Mapping Bounded Rationality', *American Psychologist,* 58(9), 697–720. URL = https://doi.org/10.1037/0003-066X.58.9.697.

Libet, Benjamin, Gleason, Curtis, Wright, Elwood & Pearl, Dennis (1983): 'Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential): The unconscious initiation of a freely voluntary act', *Brain,* 106(3): 623–642. URL = https://doi.org/10.1093/brain/106.3.623.

Nagel, Ernest (1961): *The Structure of Science: Problems in the Logic of Scientific Explanation,* New York: Harcourt, Brace & World.

Oppenheim, Paul & Putnam, Hilary (1958): 'Unity of Science as a Working Hypothesis', *Minnesota Studies in the Philosophy of Science* 2, 3–36.

Potochnik, Angela (2015): 'The diverse aims of science', *Studies in History and Philosophy of Science,* 53: 71–80. URL = https://doi.org/10.1016/j.shpsa.2015.05.008.

Putnam, Hilary (1967): 'Psychological Predicates', in William H. Capitan & Daniel Davy Merrill (eds.), *Art, Mind, and Religion*, Pittsburgh: University of Pittsburgh Press, 37–48.

Raatikainen, Panu (2010): 'Causation, Exclusion, and the Special Sciences', *Erkenntnis,* 73(3): 349–363. URL = https://doi.org/10.1007/s10670-010-9236-0.

Sutton, Richard & Barto, Andrew (2018): *Reinforcement Learning: An Introduction,* 2nd edition, Cambridge: The MIT Press.

Woodward, James (2000): 'Explanation and Invariance in the Special Sciences', *British Journal for the Philosophy of Science,* 51(2): 197–254. URL = https://doi.org/10.1093/bjps/51.2.197.