Reduction, Elimination and Radical Uninterpretability

David Roden

In this paper, I argue that the anti-reductionist thesis supports a case for the uselessness of intentional idioms in the interpretation of highly flexible, self-modifying agents that I refer to as “hyperplastic” agents. An agent is hyperplastic if it can make arbitrarily fine changes to any part of its functional or physical structure without compromising its agency or its capacity for hyperplasticity. Using Davidson’s anomalous monism (AM) as an exemplar of anti-reductionism, I argue that AM implies that no hyperplastic could use intentional psychology to predict its future intentional states or the psychological consequences of self-alterations. This is because AM implies there would be no laws allowing the hyperplastic to infer the psychological consequences of its self-alterations. This implies that no generalisations linking current to future psychological states would hold either - these could always be defeated by a self-intervention carried out by the hyperplastic. By the same token, neither hyperplastics nor human interpreters could use intentional psychology to understand the behaviour of other hyperplastics. Radical interpretation of hyperplastic agents - if such there were - would be impossible. It follows that were humans to become hyperplastic posthumans, intentional psychology would have to be instrumentally eliminated because neither the capacity nor the linguistic idiom for attributing propositional attitudes would retain predictive or hermeneutic utility.

Anti-reductionist physicalists or materialists deny that psychology can be theoretically reduced to physics but allow physics sovereignty concerning what exists. Anti-reductionist arguments vary but a common line of attack against reductionism is that psychology expresses rational or normative relationships between mental states; not causal or functional relationships of the kind expressed in theories of natural science. Thus in Sellars “Two Images” account physics and natural science tells us what exists but humans still encounter themselves in a normatively structured “space of reasons”. Donald Davidson refers to his own version of this position as "anomalous monism" (AM):

"Anomalous monism resembles materialism in its claim that all events are physical, but rejects the thesis, usually considered essential to materialism, that mental phenomena can be given purely physical explanations. Anomalous monism shows an ontological bias only in that it allows the possibility that not all events are mental, while insisting that all events are physical" (Davidson 2001: 214)

Davidson's account seeks to reconcile three claims that appear to be in tension: 1) that mental events causally interact with physical events; 2) that causal relations occur only where the events in question are covered by strict deterministic laws; 3) "that there are no strict deterministic laws on the basis of which mental events can be predicted and explained (the Anomalism of the Mental)."

Davidson aims to do this by arguing from the claim that the existence of causal relationships between events only implies that there is some true description of the relationship expressing a strict nomic relationship. The reconciliation is possible because causal relations obtain between token singular events while laws are linguistically expressed generalisations. Mental events can be causally related to one other or to non-mental events.

But, according to Davidson, causality is nomological only in that where two events are causally related, they have linguistic descriptions that express a law. It does not follow that "that every true singular statement of causality instantiates a law" (215). Thus a statement like "Helen's belief that Justin was murdered was caused by her seeing blood in the kitchen" adverts to a law like relationship between a token of blood in the kitchen

and a token belief about murder *but does not state it*. The law-like relationship, for Davidson, would have to be expressed in terms of the states and dynamics of a physical system which allowed a deterministic inference about a future state - her belief token - again rendered in some physicalistic idiom.

Claim 3) Follows, Davidson thinks, if mental states are those addressed in propositional attitude ascriptions and that such ascriptions depend holistically on overall assessments of the rationality and cognizance of agents in their world. In the space of reasons, where propositional attitudes are ascribed to persons, it is always possible to revise attributions in the interests of overall cogency. There can be no single translation scheme that pre-empts all the evidence that could be relevant to such ascriptions (222-223). Thus whereas the theories in which physical regularities are stated must be closed to allow the formulation of exceptionless laws (homonomic) the language of propositional attitude ascription is necessarily open to multiple idioms or “heteronomic” (219):

“The heteronomic character of general statements linking the mental and the physical traces back to this central role of translation in the description of all prepositional attitudes, and to the indeterminacy of translation. There are no strict psychophysical laws because of the disparate commitments of the mental and physical schemes. It is a feature of physical reality that physical change can be explained by laws that connect it with other changes and conditions physically described. It is a feature of the mental that the attribution of mental phenomena must be responsible to the background of reasons, beliefs, and intentions of the individual.”(222)

In Nagelian terms, it would be impossible to formulate true bridge laws between a reducing theory in some physical idiom and a reduced psychological theory because the intentional side the biconditional could always be revised in the light of holistic considerations irrelevant to the “physical side”. Thus type-type psychophysical reduction appears impossible. Note that an analogous result is obtainable if we view the space of reasons as structured by implicit norms irreducible to behavioral regularities.

Of course, not all accounts of reduction require bridge laws between reduced and reducing theories, or treat theories as interpreted sets of sentences. It is still open to the reductionist to argue for a different form of reduction (Bickle 1993: 222-4). It is also open to the reductionist to argue that psychology is not peculiar in being inexpressible "as sets of generalizations" - this being true of all scientific theories (226) - or in being open to extra-theoretical idioms in which to describe their contexts of application to real systems. Maybe no theory (physical or otherwise) is truly heteronomic.

However, in the argument that follows I will suppose that Davidson's anomalism is right, or, at least, that his account can be rectified in a form that is proof against neoreductionist assaults.

So let us assume that the psychological perspective in which agents have beliefs and desires and utter meaningful statements is conceptually irreducible (as Sellarsians say) to the scientific image of the world as a causal-physical system.

If so, then the possibility of a certain form of technological descendant of current humans (posthumans) *implies that intentional psychology will be instrumentally if not theoretically eliminated*.

That is, whatever its current value for humans, it could not play a similar role for the relevant class of posthuman. And this not because of any logical or ontological vices but because of it would be incapable of functioning as an idiom for interpretation and understanding among these hypothetical successors. So the anti-reductionist argument against theoretical reduction/elimination supports a metaphysical case for instrumental elimination.

The hypothetical entities in question are what I refer to in *Posthuman Life* and elsewhere as “hyperplastic agents” (Roden 2014: 101-2). An agent is hyperplastic if it can make arbitrarily fine changes to any part of its functional or physical structure without compromising either its agency or its capacity for hyperplasticity. For example, suppose a hyperplastic agent dislikes some unpleasant memories associated with the taste of milk. Whereas a merely plastic agent like ourselves might need hours of cognitive behavioral therapy to excise these, the hyperplastic simply needs to locate the neuronal ensembles and pathways associated with these memories and ensure that they are no longer linked in such a way that the memory of milk causes them to activate in turn.

Likewise, a hyperplastic would be in a position to alter any other informational or value-relevant state by physically altering the relevant brain states. Obviously, use the term “brain” broadly here to refer to those systems within the hyperplastic that are associated with “cognition”, “perception” or the “control of behaviour” in some intuitive sense of these terms. The original inspiration for the idea of the hyperplastic came from Steve Omohundro’s speculations about the goal structures of generally intelligent robots in his essay “The Basic AI Drives” (2008). We need not assume that the “brain” in question is a known biological system.

Davidson’s anti-reductionism implies token physicalism (each event that can be brought under a psychological description is identical to some physical event, since ontological physicalism is taken as a given).

So for any state in an agent with a psychological description there will be physical description of that state. For any such state there will interventions that the agent can make into the state which will produce a physically distinct successor state such that the former psychological description will no longer be true of it.

Now we can suppose that any hyperplastic agent will have an *Agenda* at a particular time. That is, it will not tinker with its internal states arbitrarily but wish to do so in ways that don’t kill it, do not undermine its capacity for hyperplasticity and that fulfill whatever desiderata are listed on the Agenda.

The interesting question (assuming Davidsonian anti-reductionism) is how the Agenda can be formulated. Can it be expressed in psychological terms (roughly, in terms of propositional attitudes or values)? If it is expressed in psychological terms, then anti-reductionism implies that for any Agent intervention at the physical level, it will not be possible to reliably infer the psychological outcome of the alteration.

This follows simply because there are no psychophysical laws. Moreover even rough generalisations over past interventions would not be much help. These might be reliable for merely plastic creatures whose basic design and structure remain fairly constant over time. But a hyperplastic agent is protean. Thus it cannot assume that the rough and ready psychophysical generalisations that have held over one phase of its existence will extend into another phase.

It follows that however a hyperplastic agent frames the Agenda *it cannot be psychologically expressible* because no reliable inferences can be drawn from future physical form to future psychology.

So if hyperplastics have Agenda’s, they would have to represent states that could be reliably inferred from facts about their physical constitution at a given time. But given Davidson's anti-reductionism, they would have little use for psychological self-description for making generalisations about their current or future actions. Suppose a hyperplastic Agent self-attributes a belief *b*. A merely plastic agent like you or me might assume generalisations along the lines of “I will continue to hold b unless I find evidence from which some contrary of b can be inferred”. But a hyperplastic agent would not be able to assume such generalisations because there could be no evidence that an auto-

intervention would not cause it to lose *b* regardless of the evidence in its favour.

So a hyperplastic agent could not use propositional attitude psychology to predict its own behaviour. Folk psychology would be equally impotent for predicting the behaviour of its fellow hyperplastics for the same reason.

If hyperplastic agents could exist and plan their self-interventions, they would have to employ an entirely different idiom to understand themselves or one another. A posthuman-making disconnection that resulted in the emergence of hyperplastics would inevitably to result in the instrumental elimination of folk psychological capacities among the population of hyperplastics, at least; since neither the capacity nor the linguistic idiom for attributing propositional attitudes would have predictive or hermeneutic utility.

This means that were humans to encounter hyperplastics, they would not be radically interpretable (in Davidson's sense) because radical interpretation depends on the principle of charity and this, again, is framed in folk psychological terms.

I conclude that if hyperplastic agents are possible, we could not understand them without abandoning the conceptual framework we currently use to understand ourselves and our conspecifics. They would be radically uninterpretable.

References

Bickle, John (1992). Mental anomaly and the new mind-brain reductionism. *Philosophy of Scienc*e 59 (2):217-30.

Davidson, D. 1984. *Inquiries into Truth and Interpretation*. Oxford: Clarendon Press.

Davidson, Donald (2001). *Essays on Actions and Event*s, Vol. 1. Oxford: Oxford University Press.

Omohundro, S. M. (2008). "The Basic AI Drives". *Frontiers in Artificial Intelligence and Applications* 171: 483

Roden, David. 2014. Posthuman Life: Philosophy at the Edge of the Human. London: Routledge.