# Rule-Following and the Irreducibility of Intentional States

Antti Heikinheimo, Jyväskylä, Finland

## 1. Reduction through Functional Definition

It is not always clear what exactly is meant when it is said that something mental is reducible to something physical. Thus, when debating about reductionism, it is important to keep in mind just which kind of reduction one is talking about. One clearly defined and plausible notion of reduction comes from Jaegwon Kim. Reducibility is often taken to be a relation between two "levels", such as the mental and the physical level. Kim argues, plausibly in my opinion, that so called bridge-laws that connect the two levels with empirical regularities, do not amount to reduction (Kim 2005, 103-5). This is because both the higher- and the lower-level phenomena need to be mentioned in a statement of a regular connection between phenomena at two different levels, whereas reduction requires an account of the higher-level phenomenon solely in terms of the lower level. I take this much to be common ground between most reductionists and non-reductionists – that it is not enough for the reductionist to establish empirical connections between the mental and the physical. He/She needs something stronger. In Kim's view this stronger requirement is:

> Conceptual connections, e.g., definitions, providing conceptual/semantic relations between the phenomena at the two levels. (Kim 2005, 108)

These conceptual connections serve as the first step of a reductive explanation, in terms of the "base" level, of the phenomenon to be reduced. The reductive explanation consists of three steps:

> Step 1 (functionalization of the target property) Property M to be reduced is given a *functional definition* of the following form: Having M $=_{def.}$ having some property or other P (in the reduction base domain) such that P performs causal task C.
>
> Step 2 (Identification of the realizers of M) Find the properties (or mechanisms) in the reduction base that perform the causal task C.
>
> Step 3 (Developing an explanatory theory) Construct a theory that explains how the realizers of M perform task C. (Kim 2005, 101-2)

On this model, then, the reduction of a higher-level property, such as being a gene, consists of (1) a functional definition, such as "being a gene = $_{def.}$ being a mechanism that encodes and transmits genetic information"; (2) finding the realizers for the causal-functional role – in this case, DNA molecules; and (3) a theory – in our case molecular biology – that explains how the realizers – the DNA molecules – fulfil this role (Kim 2005, 101). In the mind-body case, the higher-level properties in question are such as "being in a mental state S".

Although Kim's notion of reduction through functional definition is not, by any means, the only intelligible concept of reduction, I will make it the target of my following discussion on reductionism. In the end of this paper I will include a very brief comment on theory reduction and reduction through mind-body identity. There are a few things to notice about this reduction schema. First, the functional definition should, of course, be adequate to the established meaning of the higher-level concept. It is sometimes said that, because of some indefiniteness of everyday-language concepts, they can not, strictly speaking, be defined. Since this is obviously not the real issue between reductionists and non-reductionists, 'definition' here should be understood in a relaxed sense, meaning something like "rough characterization". Second, it is the attainability of the functional definition in step 1 that is essential to the philosophical issue of reductionism vs. non-reductionism. If step 1 can be completed, i.e. adequate definitions of the higher-level properties can be given through causal roles, but the reduction nevertheless fails in steps 2 and 3, the resulting position will not be non-reductionism (at least not in the usual sense of that word), but eliminativism (if there are no realizers for the roles specified)[1]. Third, the philosophical debate over reductionism (or at least the one I have in mind) concerns the *in principle* or *theoretical* attainability of the functional definitions, not their attainability in practice.

We are now in a position to see what would constitute a conclusive argument for either side in the reductionism debate. The mind-body reductionist needs to show that

> MBR[2] It is in principle possible to define mental properties, adequately to the established meaning of the concepts in question, with recourse to causal-functional roles, not using mental property concepts in the *definiens*.

The non-reductionist, respectively, needs to show that MBR is not true, i.e. that it is not possible, even in principle, to give such definitions.

According to Kim, functional definitions are not attainable for concepts of phenomenal properties, but are attainable for concepts of intentional/cognitive properties, such as believing that p or desiring that q (Kim 2005). I will argue that functional definitions are not attainable in the case of intentional properties either, that is, that MBR does not hold for intentional properties.

## 2. The Normativity Argument

My argument is based on the discussion on rule-following in Saul Kripke's *Wittgenstein on Rules and Private Language* (Kripke 1982). Kripke's question was, approximately, "what makes it the case that, in saying 'plus' and using the + symbol, I mean addition and not some other function?" His answer was, roughly, that there is nothing, no fact, short of the whole practices of attributing meanings and doing addition in the community of language-users that makes the difference between my meaning the one thing or the other. Kripke specially considers one sort of facts that might be thought to make the difference. Namely, facts about my dispositions to use the word 'plus' and the + symbol. Now these dispositions are exactly the kind of causal-functional roles that appear in Kim-style reductive explanations. Furthermore, functionally defining

[1] That is, if we have conclusive grounds for claiming that there are no realizers for the causal roles. If we have just not yet managed to find the right realizers, then, of course, we do not have to give in to eliminativism.

[2] For mind-body reductionism.

intentional states requires functionally defining meaning something instead of something else. For surely we need to be able to differentiate the contents of intentional states in order to differentiate the states themselves. And if a definition does not enable us to tell the difference between, say, believing that there is a cow in front of me and believing that there is a horse in front of me, then it is clearly not adequate to the meaning of the concept of belief. Those who think that mental content does not depend on public language might object that considerations of word meaning do not apply to intentional states. I believe that mental content does depend on public language. But even if it does not, in order to have reductive explanation, we need to be able to publicly refer to specific mental contents. So the distinction between different mental contents needs to be done in public language. Thus similar considerations apply. So let us take a look at Kripke's argument against dispositional analyses of meaning.

Kripke's main argument against dispositionalism is the normativity argument, which I will now lay out. In order to make it the case that I mean anything by a word, the meaning-determining fact needs to make the difference between right and wrong uses of the word. It needs to justify my using the word the way I use it (if I actually am using it correctly). But dispositions can not do this. If what I mean by a word was determined by the way I am disposed to use it, then whatever I say would be correct (Kripke 1982, 24). I could not mistake a cow for a horse, for if I called a cow 'horse', then that particular cow would, *for that very reason,* be included among the things I mean by 'horse'. So there would be no distinction between using a word correctly, in accordance with its meaning, and using it incorrectly. From this it follows that there would be no such thing as meaning anything by a word.

There are, of course, other candidate solutions for the rule-following problem, besides the Kripkean community view and the simple dispositional view. The most promising such solutions will not, however, help the case of reductionism, since they do not offer causal-functional analyses of meaning. I have in mind here primarily the accounts of Crispin Wright and Philip Pettit, which are, in essence, versions of the community view (see Kusch 2006, ch. 7). The reductionist needs a solution close enough to the simple dispositional view to yield functional definitions.

The lesson to be learned from the normativity argument is this: Meaning is normative. In order for a word to mean something, there must be correct and incorrect ways to use the word. Any functional definition of meaning must maintain this distinction between correctness and incorrectness. Similarly, any functional definition of intentional states must maintain the distinction between fit and misfit with actual states of affairs (in case of belief this amounts to the distinction between true and false beliefs, in case of desires, satisfied and not satisfied desires, and so on). Next I will take a brief look at some causal-functional analyses of intentional states, and how the normativity argument shows them to be defective.

## 3. Functional Analyses of Intentional States

The first functional analysis I will consider is W.V.O. Quine's behavioural semantics (Quine 1960). Quine, of course, intended his analysis to be an analysis of the meaning of sentences, for he did not believe in intentional states (see Quine 1960, 221). It is, however, quite straightforward to extend the behavioural account also to mental content. Quine's basic idea was that the (stimulus) meaning of a sentence is the set of stimuli, presented with which a language user would, if queried, affirm the sentence in question (Quine 1960, 32). So it is natural to say that the same set of stimuli constitutes the content of a belief of the language user. In other words, that he/she believes the sentence to be true. Functional definitions of other intentional states along these lines may be more complicated, but it does not matter to my argument. If the behavioural account fails in the case of belief, which is the simplest case, then there is not much hope for it in other cases either. Now it is easily seen that the normativity argument refutes the behavioural account. For the behavioural account is really nothing more than the simple dispositional account already discussed. If whatever stimulus that prompts me to affirm a sentence is counted as partly determining the meaning of the sentence, then it is not possible for me to make a mistake by affirming the sentence. So in the case of belief, all my beliefs will be true, for their contents are determined by whatever the facts happen to be when I express the beliefs. Quine, of course, tried to make room for mistakes, but even he had to acknowledge that from the behavioural account follow all kinds of indeterminacy in meaning, so that it would often have to be more or less arbitrarily decided whether someone is mistaken or uses a word in an unusual way.

Another possible source for functional definitions is a sentences-in-the-head view. According to such a view, intentional states are brain states that somehow resemble public language sentences. The most important example of such a view is Jerry Fodor's language of thought -hypothesis (Fodor 1976). There are at least two possible ways to conceive of sentences in the head. They could have content in virtue of their non-causal properties, such as some kind of isomorphism with public language sentences. Or they could have content in virtue of their role in controlling behaviour. If content of brain states is due to non-causal properties, this will not help the reductionist, for the reductionist needs causal-functional definitions. If, on the other hand, content is due to causal role in controlling behaviour, the reductionist still faces the problem of defining intentional states in terms of behaviour. And as we just saw, because of the normativity condition, that problem seems hard to solve. So it seems that sentences in the head will not be of much help to the reductionist. This, of course, is not a problem for Fodor, since he is not a reductionist.

Still another reductionist theory of mental content is teleosemantics, which purports to account for content in terms of evolutionary selection history (see e.g. Millikan 1984). But teleosemantics is a historical, not a causal-functional theory. This means that, in the teleosemantic view, content does not supervene on the totality of causally relevant facts about the present (see Dretske 2006, 75). And this rules out the possibility of causal-functional definitions of intentional states. So teleosemantics is not an option for a Kim-style reductionist. Accordingly, teleosemantics does not aim at reduction through functional definition, but reduction through identity.

## 4. Conclusion

I hope my discussion this far to have shown that there are some *a priori,* philosophical grounds to doubt the possibility of mind-body reduction through functional definition. I believe, though limitations of space prevent me from elaborating the point, that similar considerations apply against theory reduction – the view that a correct theory of the mental could in principle be derived from an all-encompassing theory of the physical – since I see no other route to theory reduction besides functional definitions of the higher-level properties. Still it might be thought that the sentences-in-the-head view, as well as teleosemantics, might facilitate reduction through mind-body identity. But I think there are difficulties for this project, too. Reduction through identity is supposed to be based on an empirical discovery to the effect that some higher-level phenomenon is in fact identical with some lower-level phenomenon, as in the case of water = $H_2O$. But the water = $H_2O$ identity rests precisely on the fact that the characteristics of water can be explained in terms of water being $H_2O$. And the normativity argument shows that similar explanation of the characteristics of intentional states in terms of brain states is not to be expected. The purpose of these remarks on theory reduction and reduction through identity has been merely to hint at the direction where I think the problems are, and they are not intended to be at all conclusive.

## Literature

Dretske, Fred 2006 “Representation, Teleosemantics, and the Problem of Self-Knowledge” in: Graham MacDonald and David Papineau (eds.), *Teleosemantics,* Oxford: Clarendon Press, 69-84.

Fodor, Jerry 1976 *Language of Thought,* Hassocks: Harvester Press.

Kim, Jaegwon 2005 *Physicalism, or Something near Enough,* Princeton: Princeton University Press.

Kripke, Saul A. 1982 *Wittgenstein on Rules and Private Language,* Cambridge: Harvard University Press.

Kusch, Martin 2006 A Sceptical Guide to Meaning and Rules, Chesham: Acumen.

Millikan, Ruth Garrett 1984 *Language, Thought, and Other Biological Categories,* Cambridge: M.I.T. Press

Quine, Willard Van Orman 1960 *Word and Object,* Cambridge: Technology Press of the M.I.T.