Non-Classical Reductionism in the Philosophy of Mind and Cognition

Christian J. Feldbacher-Escamilla[1] [0000−0001−6414−2153]

University of Cologne
Department of Philosophy
Albertus-Magnus-Platz 1, 50931 Cologne, Germany
cj.feldbacher.escamilla@uni-koeln.de
http://cjf.escamilla.academia.name

Abstract. The mind-body problem is at the core of the philosophical debate about cognition. One kind of theory on the mind-body problem is physicalism. According to a strict version, all mental objects, terms, or theories are reducible to physical ones. Hence, every physicalistic theory is also a reductionistic one. The most famous representative of reductionism in the early stages of modern philosophy of science is the philosopher Rudolf Carnap, who tried to construct a framework for reductionism in his seminal work *The Logical Structure of the World*, which was and still is highly influential for the subsequent debate on reductionism in the philosophy of mind. In this contribution, we argue for an expansion of this framework by weakening some criteria of reductionism. One advantage of such an expansion seems to be increased applicability with respect to psychological theories, e.g. the theory of extended selves, which are themselves considered relevant for cognition in the sense of increasing explanatory virtues via the concept of extended cognition.

Keywords: mind-body problem · physicalism · reductionism · extended self · extended cognition

1 Introduction

There are two kinds of mind-body problems discussed by Rudolf Carnap[1]: First, the so-called *external question* on how the elements at the basis of a system, out of which the mental and physical objects are constructed – where construction is the inverse direction of reduction –, interact with each other. Carnap considers this question as a *pseudo-problem*, because it is a question not to be answered within the limits of empirically informed science: The reason why the foundational building blocks of a theory of mind and cognition are structured as they are resists further analysis; science must take their nature for granted, as they are fundamental. Second, the so-called *internal question* of how to construct mental and physical elements given a particular basis. His answer to this question

[1] In order to allow for better accessibility for an interdisciplinary community, throughout this contribution relevant literature is mentioned at the end of each section only.

is a physicalistic one: All intersubjective mental elements, that are in Carnap's view elements of the so-called *heteropsychological domain* (e.g. statements such as 'Chloé is happy.'), are constructible out of the physical and *autopsychological* (i.e. reflective mental elements as expressed via auto-phenomenological statements such as 'I feel happy.') domain alone.

This contribution systematically reconstructs Carnap's reductionist framework, outlines its common criticisms, proposes targeted improvements to resolve these issues, and highlights a potential application for the theory of mind and cognition. For this purpose, we first discuss different kinds of physicalism (section 2); since, in our strict understanding, every form of physicalism is also some form of reductionism, we are going to discuss also different forms of reductionism; we will show that, in order to be applicable, one has to weaken the reductionistic programme non-classically (both in section 3). At the end of the paper, we are going to briefly indicate an application of a non-classically weakened reductionistic programme to the theory of the extended selves, which provides an important basis for the extended mind thesis, arguing that cognitive processes and mental states such as belief and memory are not confined to the brain (section 4).

Literature. Carnap started his epistemic and philosophy of science endeavour of providing a framework for scientific construction, reconstruction, and reduction in [3]. He expanded it with his highly influential [4, 5], and provided a new reconsideration and generalisation in [7]. The most influential contribution regarding the transfer of the philosophy of science methodology on reduction to the philosophy of mind was Ernest Nagel's [13].

2 Physicalism

In the philosophy of mind, theories are sometimes divided into dualistic and monistic theories. The set of monistic theories can be divided further with the help of different criteria: the criterion of the number of different substances, the criterion of the ontological basis, and the criterion of the different kinds of reduction. So, one may divide the set of monistic theories fully and disjointly into the set of finite monism, more precisely speaking: into one-substance monism, two-substance monism, etc., and the set of infinite monism. Or one may divide the set of monistic theories fully and disjointly into materialism and idealism. Or one may divide a subset of monistic theories into different kinds of physicalism. In the following, we are going to divide a subset of monistic theories in the last-mentioned way.

There are at least two kinds of reductions in the philosophy of mind: ontological reductions and conceptual or explanatory reductions. Theories that use ontological reductions allow us to identify all objects of one domain, e.g. the domain of psychology (dealing with mental objects and vocabulary M), with objects of another domain, e.g. the domain of physics (dealing with physical objects and vocabulary P). Such a reduction is performed in anomalous monistic theories, according to which it cannot in principle yield exceptionless laws

for predicting or explaining human thoughts and actions (*anomalism*) but must nonetheless share a physical ontology (*monism*). Theories that use conceptual or explanatory reductions allow us to translate, in a more or less strict sense, terms or explanations of one discipline, say psychology, into terms or explanations of another discipline, say physics. E.g., such a reduction was attempted by exponents of psychologism in logic who tried to reduce the vocabulary of logic to the vocabulary of psychology; many philosophers of science also attempted to provide a framework for conceptual or explanatory reductions; Carnap's *The Logical Structure of the World* is highly representative of such attempts. Below, we are going to discuss only physicalism with conceptual or explanatory reduction.

As is evident from the construction of usual artificial languages, there are at least three kinds of descriptive expressions in languages: descriptive terms, descriptive sentences and descriptive arguments. Hence, we can – in accordance with this classification – distinguish three different theses of physicalism:

Thesis 1 (Different forms of physicalistic theses).

(PT1.1) Every scientific term is reducible to physical or formal terms.
(PT1.2) Every scientific sentence is reducible to physical or formal sentences.
(PT1.3) Every scientific explanation is reducible to physical explanations.

To evaluate these physicalistic theses, one has to clarify what is meant by the terms 'reducible' and 'physical'. We will skip the discussion about the adequate meaning of 'physical' (only physics or also chemistry, biology, etc.) and explicate in the following section some traditional meanings of 'reducible'. In our investigation, we are particularly focusing on the notion of *reducing terms*.

Literature. The classification of theories of philosophy of mind up to the different theses of physicalism is performed in detail in [9]. The first two theses of conceptual or explanatory reduction are, e.g., postulated in [1].

3 Conceptual or Explanatory Reduction

According to the reductionistic programme of *Logical Positivism*, exactly the following three kinds of explanatory reductions of terms are possible:

Definition 1. A singular or general term M is reducible to a set of singular or general terms P in a theory or system T if and only if:

(RC1) M is eliminable with the help of P in T, or:
(RC2) M is partially eliminable with the help of P in T, or:
(RC3) M is connected with P by so-called rules of correspondence of T.

The technically easiest way to reduce M with the help of P is to define M with the help of P explicitly in a system or theory T. E.g., one can propose a reduction of 'green' with the help of 'the wavelength of natural light reflected' by the explicit definition 'x is green if and only if the wavelength of natural light reflected by x is between $495nm$ and $570nm$.' within physics. RC1-reductions

satisfy two very important criteria: The criterion of translatability and the criterion of conservativity. According to the criterion of translatability, a reduction of a term M with the help of other terms P allows one to translate all M-sentences to equivalent P-sentences. E.g., with the given reduction, one is able to translate the sentence 'This leaf is green.' to the – in the underlying theory – equivalent sentence 'The wavelength of natural light, reflected by this leaf, is between $495nm$ and $570nm$.'. According to the criterion of conservativity, a reduction of M with the help of P results in no new claims about P-sentences. If we reduce, e.g., the term 'green' in the way just outlined, then we make no new claims about wavelengths, etc., and therefore reduce in a conservative manner.

Reductions of kind RC2 are also conservative, but not fully translatable. A simple way of reducing M with the help of P in a conservative but not fully translatable manner is to define M with the help of P partially. E.g., one may say that someone is, with respect to an intelligence test iq_1, highly intelligent if she, when tested by iq_1, passes a particular threshold of points on iq_1. Such a reduction of 'highly intelligent' with the help of 'tested by iq_1' and 'reaches a specific amount of points of iq_1' satisfies not the criterion of translatability, because the sentence 'Chloé is highly intelligent according to the *Mensa test*.' is not translatable if Chloé was never and will never be tested by the *Mensa test*.

The problem of how to reduce dispositional terms like 'highly intelligent' as a unary general term – that is, highly intelligent not only with respect to a test, but in general –, 'frustrated', 'aggressive', and so on led in the philosophy of science and the philosophy of mind to the proposal of RC3. One possibility of connecting such dispositional terms with, e.g., behaviourally interpretable terms is to formulate bilateral reduction sentences for the terms, i.e. pairs of test- and manifestation conditions. E.g., the disposition of being aggressive may be reduced with the help of the bilateral reduction sentence 'If x is jostled by someone at time t, then x is aggressive if and only if x hits someone at t.' Such reductions satisfy neither the criterion of translatability nor the criterion of conservativity. The first criterion is not satisfied if x is never jostled. The second criterion is not satisfied, because by postulating the bilateral reduction sentence one is also claiming that x always reacts to jostling with hitting, if x is jostled by someone and hits someone at all (this is a consequence of the bilateral reduction sentence). This last sentence is a new consequence and fully in terms to which 'aggressive' is reduced; hence, the reduction is non-conservative.

Since RC1 satisfies the criteria of translatability and conservativity, RC2 satisfies only the criterion of conservativity and RC3 satisfies none of them, one may say that the term 'reduction', understood in the sense of RC1, is the narrowest one; more precisely: RC1-reductions by the help of explicit definitions are logically stronger than RC2-reductions by the help of partial definitions, and RC2 reductions are logically stronger than RC3-reductions by the help of bilateral reduction sentences. The reductionistic programme of *Logical Positivism* is more or less the history of weakening the criteria for reductions in this way.

It can be shown that in the earliest reductionistic stage of *Logical Positivism*, that is the stage in which only RC1 was accepted, the previously mentioned

physicalistic theses PT1.1–3 are ordered likewise: Every physicalist who states PT1.1 also states PT1.2; and every physicalist who states PT1.2 also states PT1.3. In the later stages such an ordering is not necessarily the case: It is possible that there are RC3-reductions of all non-physical terms to physical terms and therefore PT1.1 is true; however, PT1.3 can still be false because the rules of correspondence used in the reduction may not guarantee the validity of the newly stated explanations. In fact, the creative part that comes with such rules of correspondence could even contradict assumptions made in the previous explanations. We give an example of reductions not scaling up from the term to the explanatory level later. Nevertheless, for some applications, it might be possible to get a similar ordering of the physicalistic theses, if one expands the concept of an *explanation*, e.g., from a deductive-nomological concept of explanation to an inductive one, which will be indicated below.

Many physicalists hope to solve the mind-body problem by contributing to a reductionistic programme: Some terms of psychology can be defined explicitly with the help of physical terms, some can be defined partially, and some seem to be at least connectable with physical or behavioural terms by bilateral reduction sentences. But the physicalistic thesis PT1.1 is much stronger; to support this thesis, one has to show that not only some, but all or, if weakened a bit, at least many important psychological terms are reducible to physical ones. So, the main objection against PT1.1 is the thesis that not all, and not even the most important psychological terms, are reducible to physical terms.

Criticism of PT1.1 sometimes builds on the claim that for most relevant psychological terms there are no necessary and sufficient conditions expressible in physicalistic terms alone. In support of the opponents' thesis, especially dualists claim that even the latest reductionistic stage, that is, the stage in which also RC3 is accepted, is not satisfactory for reducing dispositional terms. In this opinion, one may be able to reduce a dispositional term partly, but not fully, to physical terms by providing some conditions of testing. One may say, e.g., that x is aggressive if the following holds: If someone jostles x at some time t, then x reacts with hitting the jostling person at t (test_1 and reaction_1); if someone laughs at x at t, then x shouts at the laughing person at t (test_2 and reaction_2); if someone wins against x at t in a friendly match, then x pushes the winning person at t (test_3 and reaction_3); etc. Many dualists claim that with such a list of test- and reaction-/manifestation-conditions, the term 'aggressive' cannot be fully reduced to physical terms like 'punches someone', 'speaks very loud', etc.; and the problem seems to be not the fact that such an exhaustive list of usual tests and reactions cannot be given – psychologists create such lists with every questionnaire they produce; such lists often contain more than hundreds of test-manifestation pairs. In the opinion of many dualists, the problem is rather that there are no strict regularities between tests and reactions. So, an aggressive person may sometimes hit someone who jostled her, but not always. E.g., the aggressive person could be too busy, too weary or too weak to hit. The list of reasons that possibly prevent an aggressive person from acting in a usual

manner is potentially infinite; and because of this reason, dualists think that psychological terms are not reducible to physical ones.

Some authors argue that the lesson one should draw is to give up the reductionistic programme altogether. But this seems to be a too hasty conclusion. Rather, we want to suggest another modification of the reductionistic programme. The idea is to think about another form of weakening the criteria that should be satisfied by reductions. We have seen that the main problem of using bilateral reduction sentences for reducing dispositional terms of psychology to physical terms lies in the fact that there are mostly no strict regularities between test- and manifestation-conditions. For this reason, psychologists use, in general, statistical laws and regularity claims rather than strict ones. Analogously, one may weaken RC3-reductions in the following way:

(RC4) M is connected with P by so-called rules of correspondence of T classically or non-classically.

In RC3, a connection of M and P is always classical, that is, expressed in classical logic. In comparison to this, the connection of M and P according to RC4 may be expressed in a non-classical logic. One type of non-classical logic, which is relevant for non-strict or so-called *ceteris paribus* regularities, is *default logic*. In default logic, one tries to express not only strict regularities, formulated with the quantifiers 'all' and 'exists', but also quasi-general claims, formulated with the terms 'normally' or 'often', etc. One of the simplest ways of reducing dispositional terms according to RC4, is to give a set of non-classical bilateral reduction sentences for the term. E.g., we may say that someone is aggressive if test_1 *normally* leads to reaction_1; test_2 *normally* leads to reaction_2; test_3 *normally* leads to reaction_3; and so on.

We have already mentioned above that it is possible to support the physicalistic thesis PT1.1 without supporting PT1.3. This is particularly the case with RC4-reductions in the light of the classical concept of an *explanation*: An argument is an explanation of the fact described by the conclusion of the argument if all premises of the argument are confirmed strict laws, boundary conditions or auxiliary claims and the argument is classically valid. To explain, e.g., the fact that Chloé acted aggressively at an instance, one may give the following argument:

1. If x is frustrated at time t, then x is aggressive at t.
2. Chloé is frustrated now.
3. Hence, Chloé is aggressive now.

The argument is classically valid, the strict law (1) is confirmed, and the boundary condition or auxiliary (2) is true (by assumption). Hence, this argument is an explanation of the fact described by the conclusion (3). But what about a reduction? A reduction of 'aggressive' has already been mentioned. The term 'frustrated' may be connected to behavioural terms, e.g., by the non-classical rule of correspondence 'Normally x is frustrated if x fails an examination.' So the explanation may be reduced to:

1'. If x fails an examination, then it holds that if x is tested at t by test_1, then x normally shows reaction_1 and if x is tested at t by test_2, then x normally shows reaction_2 and ...
2'. Chloé just failed an examination and was at the same time tested by test_2.
3'. Hence, Chloé just showed reaction_2.

As one can see, this argument is no (deductive-nomological) explanation of the fact described with the conclusion (3'): The first premise is no strict law and the argument is not classically valid; perhaps Chloé normally shows reaction_2, if tested by test_2, and hence the two premises of the argument may be true or confirmed; but this does not guarantee that the conclusion is true; possibly Chloé shows exactly just now, not reaction_2. But the argument may count as an explanation if, in fact, Chloé showed just now, reaction_2 and if the argument is, although not classically valid, valid in a non-classical sense; and in fact, this argument is in a non-classical sense valid: normally, if the premises are true, also the conclusion is true. A formal framework for such *defeasible* reasoning is provided, e.g., by *non-monotonic logic*.

So although PT1.1 may be true, PT1.3 may be false, depending on the interpretation of 'reducible' in PT1.3: if classically interpreted, then the answer is no; if non-classically interpreted, then the answer is possibly yes. To sum up: Thesis 1.1 seems to be supportable only in the case that some criteria of the reductionistic programme are weakened: we no longer ask for strict translatability, we no longer ask for strict conservativity, and we no longer ask for strict (i.e. classical) validity.

One might counter that such a weakening simply blurs the line demarcating reductionist and non-reductionist accounts. And, in fact, whether weakening the constraints for reductions is feasible will heavily depend on the purposes we put reductions to use. If the purpose at hand is about making accurate predictions, defeasible and probabilistic reasoning proves to be a successful tool. If reductionists and non-reductionists, such as dualists, agree on this, their conflict might turn out to be only a verbal dispute. For accurate prediction, it then does not matter whether one calls the method 'reductionist in a weakened sense' or 'non-reductionist in a strict sense'. If the purposes at hand, however, are to provide *real* or *metaphysical* explanations, *real* understanding or a correct grasp of the *metaphysical* situation at hand, etc., then our distinction does not provide a helpful solution to the problem. However, then the dispute seems to be also a matter of *ontological* non-/reducability. We think that such a criticism is highly feasible, and we hope that our account indicates the relevance of purpose-related thinking about this matter.

Literature. RC1 and RC2 are accepted inter alia in Carnap's *Aufbau* [3, sect. 38f]. RC3 is accepted inter alia in [4, sect. 8]. In [6, pp. 197f], Carnap mentions that he prefers more complicated RC3-reductions than bilateral reduction sentences only. In [14], Carnap's *Aufbau* is classified as anomalous monism with the main thesis that "Objects of higher level(s) cannot be (logically) reduced to the objects of lower level" [14, p. 291]. Our RC4 counters this claim by extending the

underlying logical framework ('logic' also includes *default logic*). Systematic reconstructions of Carnap's reductionism and its link to contemporary philosophy of mind and cognition debates with applications are provided in [15, 10, 11].

4 Extended, Minds, Extended Selves, and Non-Classical Reductionism

In this section, we are going to draft an RC4-reduction of components of a theory of the extended mind and the extended self. The main claim of theories of the extended self is that we tend to regard objects, which we own, as part of our self, and we express personal and social characteristics through ownership over objects and see such objects as extensions and manifestations of ourselves. More precisely, the relevant part of the main claim of theories of the extended self is:

Thesis 2 (Extended Self). The more a person x reflectively owns an entity y at t, the more x tends to regard y as part of itself at t.

Within the philosophy of mind, such an account of the extended self is assumed to help understand and explain human cognition and identity. Accounts of the extended self usually build on an account of the extended mind, which proposes:

Thesis 3 (Extended Mind). External objects such as notebooks, smartphones, or tools and gadgets can act as physical vehicles for cognitive processes like memory and calculation.

Building on this, extended self proponents state that if our cognitive processes and functional identities are constituted by external artefacts and social structures, then our very selves and personal identities are similarly extended. In this sense, extended mind theorists often state a strict implication between the extended mind thesis (thesis 3) and the extended self thesis (thesis 2).

One way such an implication between *extended minds* and the *extended selves* helps us to understand elements of the mental domain is with respect to *personal identity*. According to traditional accounts of personal identity along the lines suggested by John Locke, personal identity is a matter of *psychological continuity*, in particular a matter of the continuity of the *mind*. However, if personal identity is closely linked to the mind, and if we extend our notion of the mind (by thesis 3), then we need to also extend our notion of personal identity (by thesis 2). And it is exactly an implication between the two theses that helps us to account for this. The extended mind/extended selves reasoning is abductive and goes along the following lines:

1. Personal identity is continuity of the mind. (Lockean tradition)
2. Psychological continuity is extended. (thesis 3)
3. Hence, personal identity needs to be extended. (from 1 and 2)
4. A good/the best way to account for 3 is to establish an implication between the extended mind and the extended self (between thesis 3 and thesis 2).

Now, how can such an implication be established? The hope is that it is grounded in a physicalist reduction. What is a physicalistically reduced version of thesis 2? It seems to be plausible to interpret the term 'tends to' as an indicator for a comparative expression: The more a person x reflectively owns an entity y at t, the more x regards y as part of itself at t. Let us assume that the term 'reflectively own' is already physicalistically reduced, so to say: let us skip the reduction of this term and let us concentrate on a reduction of the psychological term 'regards ... as part of itself'. There are many bilateral reduction sentences one may formulate to connect 'regards ... as part of itself' with physicalist terms, but for the discussion of extended selves, one connection seems to be very important: The more x regards y as part of itself at t, the less x will be willing to have y alienated away at t, i.e. the more x will consider y as inalienable at t (with a broad set of implications such as the performance of defensive actions in case of alienation). With this partial reduction at hand, one may explain the fact that some external objects, tools and gadgets are considered to be less or more integral with respect to one's identity than others. For a person with brain injuries, a particular smartphone is in general less functionally relevant for their cognitive labour than, e.g., a *brain-computer interface* that improves cognitive work (analogous to a cochlear implant for auditory processing). In this sense, such an interface is more integral to the extended mind of that person and, by this, also more identity-forming. Of course, such a relation does not hold strictly; sometimes and in some contexts people might have different dependencies, even given a similar set-up as outlined here; but normally, such a relation holds – or at least it holds statistically. But then, also the relation between the extended mind and the extended self holds not strictly, but normally only. Normally, an extended self (identity integration) is implied by an extended mind (cognitive roles), but not always. For that reason, such a reduction cannot be simply a reduction of type RC3. Rather, we can only *defeasibly* infer this feature and it has to be considered an RC4-reduction of the extended self based on the extended mind. We account for the need for an extended self (3) in light of personal identity (1) and the extension of the mind (2) via a defeasible implication between an extension of the mind towards the extension of the self (4).

Literature. The first theory of the extended self is to be found in [2]. The most prominent account of it within the philosophy of science is proposed in [8], stating an implication between the extended mind and the extended self thesis: "Does the extended mind imply an extended self? It seems so." (cf.p.18). For a recent discussion cf. [12].

5 Conclusion

In the history of reductionism, the criteria for acceptable reductions have been gradually scaled back, but only classically. We have argued here that for reductionism to be applicable in the philosophy of mind and cognition, a non-classical weakening is needed, and we have outlined an application for the case of extended mind/extended selves reasoning.

Acknowledgments. This study was funded by the German Research Foundation, DFG (*Rudolf Carnap, the Problem of Induction, and the Choice of Scientific Frameworks*, grant number 545054032).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this contribution.

References

1. Beckermann, A.: What is property physicalism? pp. 152–172. Oxford University Press, Oxford. https://doi.org/10.1093/oxfordhb/9780199262618.003.0009
2. Belk, R.W.: Possessions and the Extended Self. The Journal of Consumer Research **15**(2), 139–168 (1988), http://www.jstor.org/stable/2489522
3. Carnap, R.: The Logical Structure of the World and Pseudoproblems in Philosophy. Open Court Classics, Open Court, Illinois (1928/2003)
4. Carnap, R.: Testability and Meaning. Philosophy of Science **3**(4), 419–471 (1936). https://doi.org/10.1086/286432
5. Carnap, R.: Testability and Meaning - Continued. Philosophy of Science **4**(1), 1–40 (1937). https://doi.org/10.1086/286443
6. Carnap, R.: Psychology in Physical Language. In: Ayer, A.J. (ed.) Logical positivism, pp. 165–198. Library of philosophical movements, The Free Press, Glencoe (1959)
7. Carnap, R.: Replies and Systematic Exposition. In: Schilpp, P.A. (ed.) The Philosophy of Rudolf Carnap, pp. 858–1013. Open Court, La Salle (1963)
8. Clark, A., Chalmers, D.: The extended mind. Analysis **58**(1), 7–19 (1998). https://doi.org/10.1093/analys/58.1.7
9. Crane, T.: Dualism, monism, physicalism. Mind & Society **1**, 73–85 (2000). https://doi.org/10.1007/BF02512314
10. Feldbacher-Escamilla, C.J., Sekatskaya, M.: Reductionism, supervenience, and carnap's account of empirical confirmability. Journal for General Philosophy of Science **56**(3), 345–371 (2025). https://doi.org/10.1007/s10838-025-09728-6
11. Feldbacher-Escamilla, C.J., Sekatskaya, M.: Abduction in philosophy of mind. In: Hüttemann, A., Schurz, G. (eds.) Inductive Metaphysics: Insights, Challenges, and Prospects, pp. 128–148. Routledge/Taylor & Francis Group, New York (2026). https://doi.org/10.4324/9781003514404-9
12. Heersmink, R.: Varieties of the extended self. Consciousness and Cognition **85**, 103001 (2020). https://doi.org/10.1016/j.concog.2020.103001
13. Nagel, E.: The Structure of Science. Problems in the Logic of Scientific Explanation. Harcourt, Brace & World, INC., New York (1961)
14. Nasrin, M.: Anomalous Monism in Carnap's "Aufbau". Erkenntnis (1975-) **60**(3), 283–293 (2004), http://www.jstor.org/stable/20013263
15. von Kutschera, F.: Carnap und der Physikalismus. Erkenntnis **35**(1/3), 305–323 (1991), http://www.jstor.org/stable/20012373