Journal for General Philosophy of Science (2025) 56:345–371
https://doi.org/10.1007/s10838-025-09728-6

ARTICLE

Reductionism, Supervenience, and Carnap's Account of Empirical Confirmability

Christian J. Feldbacher-Escamilla[1] · **Maria Sekatskaya**[2]

Accepted: 11 March 2025 / Published online: 5 May 2025

Abstract
Rudolf Carnap was one of the earliest proponents of logical positivism/empiricism to explicitly discuss reductionism in relation to mental phenomena from a philosophy of science perspective. In order to address early criticism, Carnap's account underwent several modifications. An important feature of the 'mental-to-physical' reduction endorsed by the later Carnap is that of empirical confirmability. According to this feature, a reduction can only succeed if it carries with it the possibility to confirm or undermine mental claims on the basis of physical claims. Although other modifications of Carnap's reductionistic account have already been linked to the modern philosophy of mind debate (von Kutschera Erkenntnis 35:305–323, 1991), the role of empirical confirmability in this context remains unexplored terrain. In this paper, we tackle this question and identify relevant modifications in the philosophy of mind with respect to physicalism. We suggest that if one shifts from the metaphysical notion of supervenience to an epistemic derivative of it, one can closely relate supervenience physicalism to Carnap's account of empirical confirmability. We will show how this both illuminates Carnap's reductionist program from a philosophy of mind perspective and highlights how the philosophy of mind debate can benefit from Carnap's account of reduction.

Keywords Reduction · Supervenience · Physicalism · Philosophy of mind · Carnap

1 Introduction

The methodological discussion of reductionism has a far-reaching history within the philosophy of science. Particularly, Rudolf Carnap was one of the first proponents of logical positivism and empiricism to explicitly discuss reductionism with respect to the mental

✉ Christian J. Feldbacher-Escamilla
cj.feldbacher.escamilla@uni-koeln.de

Maria Sekatskaya
maria.sekatskaya@gmail.com

[1] Department of Philosophy, University of Cologne, Albertus Magnus Platz 1, 50931 Cologne, Germany

[2] Department of Philosophy, University of Duesseldorf, Werdener Str. 4, 40227 Duesseldorf, Germany

from a philosophy of science perspective as early as in his (1928/2003). Reductionism also plays a central role in debates within the philosophy of mind. However, the predominant notion used there stems from Ernest Nagel (1961), a notion that was also proposed by the early Carnap but then increasingly modified and adapted in order to keep up with new developments in science as well as its logical methodology. It seems that earlier reductive accounts were rooted not in Nagel (1961) but rather in a broader philosophical debate about reductionism and the relationship between different scientific disciplines, such as biology, psychology, and physics, oftentimes without providing references to a particular debate.

In this paper, we want to investigate Carnap's development with respect to reductionism and link it to debates and accounts of the philosophy of mind. Surprisingly, Carnap's reductionism did not gain much attention within this field. Although, e.g., van Riel (2014, 154) and van Riel and Gulick (2019, sect.1) speak about the 'Carnap-model' of reductionism, they refer only to his early account of explicit definability. Also, quite interestingly, the most influential account of Nagel (1961) does not mention Carnap in the context of reduction. Although recently Leitgeb and Carus (2020, Supplement E, section 1) provided a comprehensive overview of Carnapean accounts of reduction (they cover basically all accounts of reduction we will also cover), it were mainly (Heidelberger 2003; Kim 2003), and already (von Kutschera 1991) who discussed Carnap's reductionistic development and linked it to debates within the philosophy of mind. With a focus on Carnap, it is particularly (von Kutschera 1991) who provides the most comprehensive view. However, Franz von Kutschera's investigation is in parts quite sketchy and also has a gap with respect to what we call the *empirical confirmability account of reduction*, which, as we will argue, can be connected to models of supervenience physicalism with an epistemic reading of supervenience. Also, von Kutschera highlights more the "linguistic-pragmatic" turn of Carnap (vs the traditional focus on ontology; cf. 306), whereas our contribution stresses more the development of Carnap's account in terms of progressively weakening constraints, culminating in his concept of reduction as empirical confirmability. Despite these differences, it is important to mention that von Kutschera's work is seminal for our investigation and, most probably due to its availability in German only, is too often neglected in discussions of reductionism (e.g., *Springerlink* lists only 3 citations, 2025-01-01).

Although the main aim of our investigation is to outline Carnap's reductionistic development and indicate which models and accounts of the philosophy of mind align well with different stages of this development, we also want to indicate a consequence one can draw from the more historical discussion for the systematic framing of positions within the philosophy of mind. Namely, that the notion of *reduction* should not be viewed as a binary property that distinguishes reductive from non-reductive accounts; rather, it should be understood as a gradual notion, distinguishing stronger from weaker forms of reductionism. Our general historio-systematic hypothesis about better linking debates within the philosophy of mind and the philosophy of science is that, if debates within the philosophy of mind had not been stuck at Nagelian reduction but had continued to follow the philosophy of science development of reductionism such as that of Carnap, this gradual vs binary notion of reductive accounts would have become clearer earlier on.

Our investigation proceeds as follows: In section 2, we outline the big picture of philosophy of mind, a set of relevant conditions of adequacy, and accounts that are relevant for our investigation. In section 3, we discuss Carnap's reductionistic development. In section 4, we link the relevant accounts of the philosophy of mind to Carnap's reductionistic accounts and discuss how the latter fare with respect to the conditions of adequacy from the philosophy of mind. We conclude in section 5 with a short comparison of the adequacy

profiles of the accounts and an outline of what one might take away from this investigation for the framing of accounts within the philosophy of mind.

2 Philosophy of Mind and (Non-)Reductionism

In this section, we provide a very rough and general overview of reductive and non-reductive accounts in the philosophy of mind and pick out particular accounts that, as we will argue in the subsequent sections, serve as philosophy of mind models of the different approaches to reduction as proposed by Carnap.

We can sort the landscape of positions within the philosophy of mind along two dimensions. On one dimension, we distinguish between physicalist and non-physicalist accounts. Physicalists state that all (actual) objects are physical, whereas non-physicalists state that not all (actual) objects are physical. On the other dimension, we distinguish reductivist and non-reductivist accounts. Reductivists claim that not only all objects (tokens) are of one kind, but also that there is only one kind of property (i.e., all properties are of the same type). Non-reductivists claim that some properties of objects (tokens) are irreducible to properties of a certain type (in particular, mental properties, considered as a type 'mental,' are irreducible to physical properties, considered as a type 'physical'). If we square the two dimensions, we get four positions: (1) reductive physicalism, also 'type identity theory', stating that all objects and properties are physical; (2) non-reductive physicalism, also 'token identity theory', stating that all actual objects are physical but that some of their properties, namely mental properties, are irreducible to the physical properties; (3) reductive non-physicalism, also 'idealism', stating that all objects and properties are non-physical; and (4) non-reductive non-physicalism, also 'dualism', stating that not all objects and properties are physical (but some are also irreducibly mental).

Regarding (2), it is often assumed in the debate that there is a distinction between non-reductive physicalism and property dualism. While both views maintain that mental properties are irreducible to physical properties, they differ in the ontological conclusions derived from that. According to the property dualist, mental properties are irreducible to physical properties and cannot be identical to them due to their different modal implications, even if they happen to be nomologically co-instantiated because of supervenience. This can be framed in terms of the (metaphysical) possibility of zombies in nomologically different possible worlds, or in terms of different conditions of persistence for physical substrates and their supervenient mental properties (cf. Stoljar 2010; Schneider 2012). Non-reductive physicalists do not go as far as to conclude the ontological non-identity of properties from irreducibility. For example, one can be a functionalist regarding what is the proper analysis of mental terms but an identity theorist regarding the ontology of each particular realization (realization physicalism). Whether one can be a non-reductive physicalist (for example, a role-functionalist or, in general, a non-reductive physicalist who explains supervenience as a relation that is neither identity nor realization, as we will discuss later in the paper) without being a property dualist is a matter of contention (cf. Schneider 2012; van Riel 2014; Hüttemann 2023). We do not want to address this issue directly at this point, but we hope that by clarifying different versions of reduction, our paper will help bring some clarity to this debate (we thank an anonymous reviewer for drawing our attention to this issue).

In this paper, we focus on physicalist positions of the philosophy of mind, which is why we only further discuss non-reductive physicalist accounts. These accounts allow

for kind-distinctions on the property, i.e. type, level. However, according to Jaegwon Kim, all physicalists share the commitment of at least the "supervenience of mental on the physical", which he labels "minimal physicalism" (cf. Kim 2005, 13). That the mental supervenes on the physical can be formulated as a no-difference claim: that there is no mental distinction of objects without there being a physical distinction. We will provide a more specific characterisation of supervenience in section 3. However, for drawing a general picture of non-reductive physicalist accounts in the philosophy of mind, it suffices to note that supervenience comes with different modal strength: (2.1) realisationists assert at least *nomological supervenience* (the no-difference claim holds with nomological necessity); (2.2) necessitationists assert *metaphysical supervenience* (the no-difference claim holds with metaphysical necessity); and (2.3) at least some grounding accounts assert *logical supervenience* (the no-difference claim holds analytically). So, we end up with a rough picture of philosophy of mind positions as provided in Fig. 1.

The accounts outlined in this rough scheme arose in order to address problems of physicalist views in the philosophy of mind. The following classical problems are most relevant for our discussion—we present them as conditions of adequacy that are often seen as necessary or at least strongly desirable for any account of the mental:

- $\mathcal{A}_{\mathcal{A}}$ *Autonomy* of the special sciences: The special sciences are to a high degree autonomous from foundational sciences like physics. (cf. Fodor 1974)
- $\mathcal{A}_{\mathcal{C}}$ Mental *causation*: Some mental states are causally efficacious. (cf. Kim 1989)
- $\mathcal{A}_{\mathcal{K}}$ Natural *kinds*: Some mental properties are natural kinds. (cf. Fodor 1974)
- $\mathcal{A}_{\mathcal{M}}$ *Multiple* realisability: One and the same mental state can be realised in multiple ways. (cf. Putnam 1967)
- $\mathcal{A}_{\mathcal{R}}$ Mental *residue*: If a reduction of the mental succeeds, then there is no mental residue such as *qualia* etc. (cf. Kim 2005)

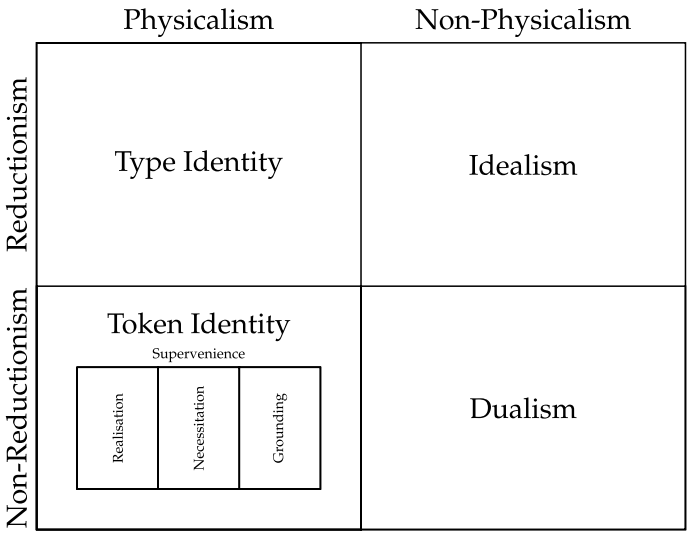

Fig. 1 Positions in the philosophy of mind: the big picture

Now, as we will see below in our discussion of Carnap's accounts of reduction and their connection to accounts within the philosophy of mind, particularly relevant are *type identity theory*, *functionalism* as a species of realisationist accounts, and *supervenience physicalism*. For this purpose, we provide a very short sketch of them here (we present only the very rough ideas and how they fare with respect to the conditions of adequacy; relevant details will be provided in section 4, when we link the philosophy of mind approaches to Carnap's accounts of reduction).

Let us begin with type identity theory, outlined in classical texts by Place (1956), Feigl (1958b), and Smart (1959). According to proponents of this theory, mental states and processes are identical to physical states and processes, even if we may not currently be able to define them in physical terms. This provides an ontological reduction of types: the claim that mental properties are identical to physical properties. Of course, type identity theorists would also agree with a weaker ontological claim that each instantiation of a mental property is identical to an instantiation of a physical property (ontological reduction of tokens). Type identity theory was criticised by Putnam (1967) and Fodor (1974) primarily because it failed to account for multiple realisability, i.e. for a plausible thesis that qualitatively identical mental states can be realised by very different physical states. For instance, pain in humans can be realised by C-fibers firing, but in other species like octopi—or even extraterrestrial life forms—the physical basis for pain may differ. If these varying physical realisations belong to different physical kinds, they cannot all be identical to the single mental kind referred to as 'pain' (the received view seems to be that this criticism serves as a decisive refutation of type-identity theory (cf. Kim 2005; Michel 2023); however, for a more type-identity-theory-friendly analysis, see Polger (2011)). Consequently, the identity theory fails to meet $\mathcal{A}_{\mathcal{M}}$ (multiple) because identification as a one-to-one correspondence excludes a many-to-one correspondence, and $\mathcal{A}_{\mathcal{A}}$ (autonomy) since all objects and properties fall squarely within the realm of physics. Even if one tries to account for $\mathcal{A}_{\mathcal{M}}$ by providing a disjunctive analysis (cf. Clapp 2001), i.e. by combining all multiple realisers disjunctively in order to avoid a many-to-one identity, one still gets in conflict with $\mathcal{A}_{\mathcal{K}}$ (kinds) because such disjunctions are too heterogeneous to form natural kinds. Whether or not this problem is a reason to reject disjunctive approaches is a matter of ongoing debate, and we refrain from taking a position here (cf. Clapp 2001; Walter 2006).

However, identity theory fares well with $\mathcal{A}_{\mathcal{C}}$ (causal exclusion), because by identifying the mental with the physical it avoids the causal exclusion problem. $\mathcal{A}_{\mathcal{R}}$ (residue) can also be satisfied, because the identification of qualitative mental states with certain physical states does not require that these mental states be analysed in physical terms. Instead, this identification is inferred as the best explanation of empirical data, in particular, of constant correlations between all observed instances of mental and physical phenomena (primary texts: Place (1956), Smart (1959)). For further analysis along these lines, see Polger (2011) and Michel (2023). In section 4, we will demonstrate that this type of identification is also what Feigl (1958b) had in mind, contrary to the interpretation of von Kutschera (1991), which focuses solely on the linguistic-pragmatic turn.

Let us come to *realisation* accounts. We focus on two forms here: logical behaviourism and functionalism. The first form is logical behaviourism, proposed by Ryle (1949). According to Ryle, all mental phenomena can be defined in terms of either publicly observable behaviours of agents or dispositions of agents to produce publicly observable behaviours in certain types of situations. Ryle proposed plausible dispositional analyses of some mental terms such as "knowledge" and "belief", and argued that any mental terms referring to real psychological phenomena could be analysed in this manner. Terms that cannot be so analysed, he contended, do not refer to anything real. They are remnants of an outdated

dualist theory of mind and should be eliminated from psychology, just as phlogiston was eliminated from chemistry.

Logical behaviourism meets three of the conditions of adequacy well. $\mathcal{A}_{\mathcal{M}}$ (multiple) is unproblematic because the same dispositions can be realised by very different physical systems. $\mathcal{A}_{\mathcal{K}}$ (kinds) is also unproblematic because even though Ryle himself did not say anything about mental kinds being natural, the specific problem of identifying one (natural) mental kind with many heterogeneous physical kinds does not arise for him. Logical behaviourism is consistent with the claim that one mental kind is realised by different physical kinds which have the same causal role in particularly defined situations. $\mathcal{A}_{\mathcal{C}}$ (causal) is satisfied because by introducing dispositions Ryle effectively identifies each mental property with the causal role it plays. $\mathcal{A}_{\mathcal{A}}$ (autonomy) is not satisfied because logical behaviourism aims for a definition of all mental phenomena in physical terms: once such a definition is achieved, mental terms can be eliminated. For Ryle himself, though, this is presumably not a problem, but rather an achievement. The biggest problem, however, is that $\mathcal{A}_{\mathcal{R}}$ (residue) is not satisfied at all. It turns out that not only qualia, which resist functional definability even according to the majority of the later functionalist accounts, but even more straightforwardly "cognitive" mental states cannot be defined purely dispositionally. For example, it seems undeniable that a paralysed person uncapable of any behaviours can still have mental states, or that a highly skilled actor can display behaviours which do not correspond to his mental states, even if this paralysed person never regains control of her body and even if this actor's pretending is never revealed as untrue.

The insights of logical behaviourism and its problems led to the development of different functionalist accounts, starting from Putnam's machine state functionalism (1960, 1967), to Lewis's (1966; 1972; 1980; 1994) analytic functionalism, which defines mental terms by means of their causal roles, to contemporary versions of functionalism, which keep the causal role definitions of the mental and differ regarding how we identify these causal roles (purely by following our folk-psychological theories of mind or also by incorporating the insights from cognitive psychology and neuroscience?). They also differ in their ontological commitments (realization physicalism vs role physicalism) and in the specific details of functional analyses of phenomenal concepts. Due to the fact that there are many different versions of functionalism with their specific strategies, we cannot go into more details here (for a recent overview, cf. Levin 2022).

However, we can evaluate how well these functionalist accounts, which all share a common core, meet the adequacy criteria. $\mathcal{A}_{\mathcal{A}}$ (autonomy) is satisfied because higher-order sciences are necessary to establish important causal correlations—such as 'pain causes avoidance reactions,' 'frustration causes aggression,' or 'rational agents aim to maximise their utility functions'—and make corresponding empirically testable predictions. $\mathcal{A}_{\mathcal{M}}$ (multiple) is unproblematic, because each functional term is defined by a particular causal role that different physical processes and states play in different physical systems. $\mathcal{A}_{\mathcal{C}}$ (causal) is also unproblematic, because mental terms are explicitly equated with their causal roles: to be a mental state or process means to have a certain causal role. However, this problem arises for role-functionalists who want to claim that functional states themselves, in virtue of being the states defined by their causal roles, have a causal influence over the outcomes in such a way that this influence is not reducible to the causal influence of the physical realisers of these functional states (cf. McLaughlin 2006; Moore 2011). In our overview, we focus on realiser-functionalists, who do not have this problem, in particular on David K. Lewis, whose account we will analyse in more detail in subsection 4.2. $\mathcal{A}_{\mathcal{K}}$ (kinds) is satisfied: the issue of identifying one mental kind with many heterogeneous physical kinds doesn't come up. This is because functionalists typically don't claim that

functionally defined types, which satisfy the requirements for natural kinds, are identical to some physical types. However, $\mathcal{A}_{\mathcal{R}}$ remains problematic. Some of the most famous debates in the philosophy of mind center around the (im)possibility of providing a functional analysis for specific types of mental terms, particularly those that refer to qualitative phenomenal experience. For different functionalist answers to this problem, cf. Shoemaker (1975), Lycan (1987), and Levin (2002); for a pessimistic conclusion, see Kim (2005).

Finally, let us come to supervenience physicalism, i.e. the versions of non-reductive physicalism that accept minimal physicalism and the thesis that mental states supervene on physical states but do not accept the further claim that supervenience holds either because mental states are identical to or realised by the physical states. According to the supervenience physicalists, a supervenience relation holds due to some form of metaphysical or nomological necessity. This implies that while in the actual world all mental properties supervene on physical properties, in other possible worlds they could exist without their physical base (cf. Stoljar 2010). Recently, the metaphysical necessitation relation between the physical and the mental has been interpreted as a form of grounding (cf. O'Conaill 2018). Since these supervenience physicalist accounts are quite diverse—and some are also quite new—it is difficult to characterise them all in a short and unified manner. However, some observations can still be made regarding their conditions of adequacy.

$\mathcal{A}_{\mathcal{M}}$ (multiple) and $\mathcal{A}_{\mathcal{K}}$ (kinds) are unproblematic, because each particular mental property that constitutes a natural kind can supervene on different physical bases belonging to heterogeneous physical kinds. $\mathcal{A}_{\mathcal{A}}$ (autonomy) is satisfied because mental terms are explanatorily indispensable: the only way we can know about the mental (at least about the phenomenal) is by some sort of first-person acquaintance. $\mathcal{A}_{\mathcal{R}}$ (residue) is also unproblematic: we do not aim at providing an analysis of all mental terms in non-mental terms. If some mental phenomena, such as qualia, cannot be functionally analysed or identified with some particular physical states, they can still supervene on some physical states. However, $\mathcal{A}_{\mathcal{C}}$ (causal) is notoriously problematic for supervenience physicalists due to the causal exclusion argument which shows that if mental properties supervene on the physical properties but are neither identical to them nor can be defined in terms of their causal roles, and if all physical effects have physical causes (the principle of the causal closure of the physical), then mental properties are causally inert (cf. Kim 2005). There are attempts to answer the causal exclusion argument (cf. List and Peter 2009; Kroedel and Moritz 2016; for criticism, cf. Gebharter 2017), but the general agreement seems to be that this is the most serious problem for all versions of supervenience physicalism.

We can summarise the result of our discussion as provided in Table 1.

Now, we saw that the positions in the philosophy of mind are differentiated also according to their stance with respect to reductionism. The notion of 'reduction' predominantly used in the philosophy of mind since the 1960s originates from Nagel. Nagel's general characterisation is as follows:

Table 1 Overview of the selected philosophy of mind accounts and how they meet the conditions of adequacy

Condition of adequacy		Type ident.	Disjunct.	Functional.	Sup. phys.
$\mathcal{A}_{\mathcal{A}}$	(autonomy)	−	−	+	+
$\mathcal{A}_{\mathcal{C}}$	(causal excl.)	+	+	+	−
$\mathcal{A}_{\mathcal{K}}$	(kind)	−	−	+	+
$\mathcal{A}_{\mathcal{M}}$	(multiple)	−	+	+	+
$\mathcal{A}_{\mathcal{R}}$	(residue)	+	+	−	+

> "A reduction is effected when the experimental laws of the secondary science [...] are shown to be the logical consequences of the theoretical assumptions (inclusive of the coordinating definitions) of the primary science." (cf. Nagel 1961, 352)

So, if $\mathcal{T}_{\mathcal{M}}$ comprises all laws about, e.g., the mental, and $\mathcal{T}_{\mathcal{P}}$ comprises all laws about, e.g., the physical, then, according to Nagel, we can reduce $\mathcal{T}_{\mathcal{M}}$ to $\mathcal{T}_{\mathcal{P}}$ if we show that $\mathcal{T}_{\mathcal{M}}$ follows from $\mathcal{T}_{\mathcal{P}}$ and a set of coordinating definitions $\mathcal{D}$. With respect to the nature of $\mathcal{D}$, Nagel has three possibilities in mind (cf. 1961, 354):

- $\mathcal{D}1$ $\mathcal{D}$ as providing an identification or analysis of meaning, e.g. via an explicit definition with intensional equivalence of $\mathcal{T}_{\mathcal{M}}$ terms on the basis of $\mathcal{T}_{\mathcal{P}}$ terms
- $\mathcal{D}2$ $\mathcal{D}$ as providing purely conventional definitions of $\mathcal{T}_{\mathcal{M}}$ terms on the basis of $\mathcal{T}_{\mathcal{P}}$ terms
- $\mathcal{D}3$ $\mathcal{D}$ as providing empirically justified sufficient (possibly also necessary) conditions for the application of $\mathcal{T}_{\mathcal{M}}$ terms on the basis of $\mathcal{T}_{\mathcal{P}}$ terms (the idea being that the meanings of the $\mathcal{T}_{\mathcal{M}}$ terms and that of the $\mathcal{T}_{\mathcal{P}}$ terms are independently given and then it is empirically shown that they are coextensional).

We can see that, although Nagel differentiates three kinds of coordinating definitions ($\mathcal{D}1$ –$\mathcal{D}3$), all of them ask for establishing an equivalence between the terms of $\mathcal{T}_{\mathcal{M}}$ and the terms of $\mathcal{T}_{\mathcal{P}}$. The equivalences come with different modal force: $\mathcal{D}3$ is only about coextensionality, hence, comes with no modal (intensional) force. $\mathcal{D}2$ comes with conventional modal force; and $\mathcal{D}1$ comes with analytic modal force. Still, extensionally speaking they all ask for the possibility of explicit definability (which is guaranteed by coextensionality). Since here 'reductive' means 'Nagel reductive', and 'Nagel reductive' implies explicit definability, we can observe that in general, reductive accounts in the philosophy of mind are those that require the explicit definability of $\mathcal{T}_{\mathcal{M}}$ terms based on $\mathcal{T}_{\mathcal{P}}$ terms (Hüttemann 2023; Michel 2023; van Riel 2014). Note that our discussion of Nagel's account is based on a linguistic/epistemic interpretation of Nagelean reduction. Whether an ontological (vs theoretical—cf. van Riel 2014, 19) interpretation would also be suitable, and how such an interpretation could be embedded into our framework, is currently unclear.

Having outlined relevant positions of the philosophy of mind and their location in the broader context of (non)-reductionism, we can now move on to the philosophy of science and outline Carnap's accounts of reduction.

3 Carnap on Reduction

In this section, we spell out different accounts of 'reduction' as proposed by Carnap, who started with the suggestion of a very strict programme that was increasingly weakened. Once we have worked out his development, we will interpret it in the next section on the basis of the philosophy of mind debate centering around reductionism, as outlined in the previous section.

The first account of reduction can be found as early as in his *Aufbau* from 1928 (cf. Carnap 1928/2003). There he outlined and applied a so-called "constructional system of concepts". He suggested the following understanding of reduction (in the German original, Carnap often uses the terms 'zurückführbar' for 'reducible'):

> "An object (or concept) [a] is said to be reducible to one or more other objects [(or concepts) b and c] if all statements about it can be transformed into statements

about these other objects [(or concepts)]." (p.6)
The transformation of statements has to be "into a coextensive propositional function in which *a* no longer occurs, but only *b* and *c*." (p.61)
"In the simplest case, such a translation rule will consist in the prescription to replace *a* in all its occurrences by a certain expression in which only *b* and *c* occur ("explicit" definition)." (p.61)
Synonymously for 'construction' (in the German original: 'Konstitution'):
"[T]he construction of an object must be given in the logical form of a definition: every object to be constructed will be introduced through its constructional definition[.]" (Carnap 1928/2003, 12)

Although Carnap suggested to apply this account of reduction also to psychology and statements about the mental already in the *Aufbau*, he did so more explicitly in his (1931) and (1932):

"The application of our principles to *Psychology* usually provokes violent opposition. In this department of science, our thesis takes the form of the assertion that [...] the definition of any psychological term reduces it to physical terms." (Carnap 1995, 71f; German original in Carnap 1931, 450)
"Our thesis thus states that a definition may be constructed for every psychological concept (i.e., expression) which directly or indirectly derives that concept from physical concepts." (Carnap 1959, 40; German original in Carnap 1932, 109)

So, we can say that Carnap's first account of reduction was one of explicit definability. To have a standard scheme for this:

Definition *(reduction as explicit definability) A mental predicate M is $\mathcal{C}$1-reducible to a set of physical predicates $\mathcal{P}$ iff M is explicitly definable on the basis of $\mathcal{P}$.*

Quine (1951, 36) ascribed this form of reduction to Carnap in his famous *Two Dogmas* as a form of "*radical reductionism*[:] Every meaningful statement is held to be translatable into a statement (true or false) about immediate experience [...] Carnap embarked on this project in the *Aufbau*."

$\mathcal{C}$1-reducibility comes with three important positive features: *non-creativity*, *eliminability*, and *extended reducibility*. To briefly explain these three features, let us assume that we start with a mental theory $\mathcal{T}_{\mathcal{M}}$, systematising all our scientific knowledge of the mental, and a physical theory $\mathcal{T}_{\mathcal{P}}$, and that we are able to provide a $\mathcal{C}$1-reduction of all mental predicates in $\mathcal{T}_{\mathcal{M}}$ on the basis of the predicates of $\mathcal{T}_{\mathcal{P}}$. As for *non-creativity*, we can observe that extending $\mathcal{T}_{\mathcal{P}}$ by the definitions of the $\mathcal{C}$1-reduction in order to achieve a theory equivalent to $\mathcal{T}_{\mathcal{M}}$ does not change anything with respect to $\mathcal{T}_{\mathcal{P}}$, i.e. all theorems and non-theorems of $\mathcal{T}_{\mathcal{P}}$ in terms of the predicates of $\mathcal{T}_{\mathcal{P}}$ remain unchanged. In this sense, the $\mathcal{C}$1-reduction (we could also speak of a $\mathcal{C}$1-construction) of $\mathcal{T}_{\mathcal{M}}$ on the basis of $\mathcal{T}_{\mathcal{P}}$ is conservative.

As for *eliminability*: explicit definitions allow for equivalent replacement of the definiendum (here a predicate of $\mathcal{T}_{\mathcal{M}}$) by the respective definiens (here an open sentence or formula of $\mathcal{T}_{\mathcal{P}}$). This means that, given such a $\mathcal{C}$1-reduction, whatever can be expressed by using a mental predicate of $\mathcal{T}_{\mathcal{M}}$, can be equivalently (given the definitions of the $\mathcal{C}$1-reduction) expressed using the physical predicates of $\mathcal{T}_{\mathcal{P}}$. In this sense mental predicates can be eliminated.

Finally, as for *extended reducibility*: Until now, we were speaking of $\mathcal{C}1$-reducibility with respect to predicates only. However, given the features of non-creativity and eliminability of explicit definitions, we can also extend our parlance to the $\mathcal{C}1$-reduction of laws and theories: by non-creativity and eliminability we know that all sentences or formulæ of $\mathcal{T}_{\mathcal{M}}$ can also be replaced by equivalent sentences or formulæ of $\mathcal{T}_{\mathcal{P}}$, so we gain also a $\mathcal{C}1$-reduction of the laws of $\mathcal{T}_{\mathcal{M}}$ on the basis of $\mathcal{T}_{\mathcal{M}}$. And since it is not only the laws of $\mathcal{T}_{\mathcal{M}}$ we can $\mathcal{C}1$-reduce in such a way, but any axiom (laws, auxiliaries, etc.) of $\mathcal{T}_{\mathcal{M}}$, we even gain a $\mathcal{C}1$-reduction of $\mathcal{T}_{\mathcal{M}}$ as a whole on the basis of $\mathcal{T}_{\mathcal{P}}$. So, non-creativity and eliminability allow us to extend the property of $\mathcal{C}1$-reducibility from predicates to laws and theories.

Now, these theoretical features of $\mathcal{C}1$-reducibility come at a high cost: In order to provide such a reduction one needs to find individually necessary and jointly sufficient conditions for the predicate (of $\mathcal{T}_{\mathcal{M}}$) to be reduced by the help of predicates (of $\mathcal{T}_{\mathcal{P}}$) of the reduction basis. As we saw above, with respect to such a reduction of mental predicates, a common objection stems from the literature on *multiple realisability*. Carnap responded to an earlier criticism concerning an adequate treatment of dispositional terms. If we take, e.g., the term 'soluble', then "such disposition-terms cannot be defined by means of the terms by which [...] conditions and reactions are described" (Carnap 1936, 440), simply because this would amount to a $\mathcal{C}1$-reduction that is too strong. If we take, e.g., as test condition that an object is put into water, and as reaction or manifestation condition that an object dissolves, defining the dispositional term 'soluble (in water)' by an explicit definition of the form that x is soluble (in water) iff x dissolves when put into water leads to the right result for any object that undergoes the test condition; but it leads to the wrong result for many objects that never undergo the test condition (e.g., a certain match which I completely burnt yesterday, never undergoes the test condition and so the definiens is vacuously satisfied, hence the solubility disposition would be attributed; cf. Carnap 1936, 440; this problem was also called the 'paradox of vacuous applicability', cf. Pap 1958, 210). So, dispositional predicates are one type of predicates that (often) do not allow for a satisfying analysis in terms of necessary and sufficient conditions and, hence, fail to be $\mathcal{C}1$-reducible.

Due to this problem of vacuous applicability, Carnap suggested to weaken the constraints for reduction. He suggested that dispositional predicates "can be introduced [i.e. reduced or constructed] by sentences of another form", namely *reduction sentences* (cf. Carnap 1936, 440). Expressed in our schema, reduction sentences have the following form (cf. Carnap 1936, 441; to stick to our scheme, we suppose that M is a predicate of $\mathcal{T}_{\mathcal{M}}$ and the P_i are predicates of $\mathcal{P}$, i.e. a set of predicates of $\mathcal{T}_{\mathcal{P}}$):

$$\underbrace{P_1(x,t)}_{TEST} \rightarrow \Big(\underbrace{M(x)}_{DISPOSITION} \leftrightarrow \underbrace{P_2(x,t)}_{REACTION} \Big)$$

So, e.g., if a piece of sugar is put into water at some sequence of points in time (test condition), then we assign the disposition of solubility (in water) to it iff it also dissolves within the sequence of points in time (reaction condition). This holds independently of whether we ever perform such a test or not, so, e.g., this disposition would not be attributed to a certain match which I completely burnt yesterday.

Expressed in our schema, we can say that Carnap's second account of reduction was one of bilateral reducibility:

Definition *(reduction as bilateral reducibility) A mental predicate M is $\mathcal{C}2$-reducible to a set of physical predicates $\mathcal{P}$ iff M can be linked to $\mathcal{P}$ by the help of a bilateral reduction sentence.*

Note that $\mathcal{C}2$-reducibility is more general than $\mathcal{C}1$-reducibility because any explicit definition of some mental predicate M on the basis of a set of physical predicates $\mathcal{P}$ can be transformed into a bilateral reduction sentence with a trivial antecedens and a trivial addition in the definiens.

$\mathcal{C}2$-reducibility comes with (modified) important positive features: *favourable creativity*, *partial eliminability*, and *evidential conclusiveness* (hints of the last feature can be found already in Carnap 1934/2001, par.82; but are given in more detail in Carnap 1956; and are also discussed in Hempel 1958, sect.8; for a modern discussion of the first two features cf. Schurz 2014). As for *favourable creativity*, one can see that in the scheme of reduction sentences as provided above, the time-variable (t) is not part of the attribution of the disposition ($M(x)$). For this reason, bilateral reduction sentences do not simply have the form of conditional definitions, which are non-creative (although only partially eliminable). Rather, they are creative inasmuch as they introduce new empirical consequences. However, this introduction is "favourable" because it is trackable and expresses a systematic feature of the test and reaction condition: It is simply the claim that if an object undergoes the test condition at some point in time and reacts in a particular way, it will react in the same way at all times when it undergoes the test condition.

As for *partial eliminability*, $\mathcal{C}2$-reductions allow to eliminate mental terms (of $\mathcal{T}_{\mathcal{M}}$) only in case where we can apply the test condition (and gain relevant information about the reaction). For all other cases the attribution of a mental property remains undecided. Eliminability can be increased if we provide further test-reaction pairs (in terms of $\mathcal{T}_{\mathcal{P}}$) for mental concepts (of $\mathcal{T}_{\mathcal{M}}$), simply because more cases will allow us to apply some test condition. Such a use of multiple test-reaction pairs to increase the range of eliminability of $\mathcal{C}2$-reductions comes with further creative empirical consequences (in terms of $\mathcal{T}_{\mathcal{P}}$). However, these are again trackable and systematically cross-link test and reaction conditions (e.g. that "any object that shows a positive response under the first test condition will, when put into the second test condition, show a positive response as well" etc.; cf. Hempel 1958, 71).

As for *evidential conclusiveness* Carnap (1956, 69) and Hempel (1958, 58) stress that characteristic for bilateral reduction is that it "permit[s] a set of observational data (such as '$[P_1(x,t)]$' and '$[\neg P_2(x,t)]$' above) to constitute conclusive evidence for or against the applicability of the theoretical term in a given situation." So, if we apply the test condition and check for the reaction, we will be able to in fact attribute the dispositional property or not.

Disadvantages of $\mathcal{C}2$-reducibility are that it is hard to achieve completeness because one needs to find many adequate test reaction pairs in order to increase eliminability and that we loose extended reducibility because we lack a complete term-by-term reduction (in the sense of $\mathcal{C}1$-reducibility), so we also cannot provide sentence-by-sentence reductions, i.e. in particular no complete reduction (again, in the sense of $\mathcal{C}1$) of laws and theories.

Now, there is another disadvantage of $\mathcal{C}2$-reducibility. Although it is a weakening of $\mathcal{C}1$-reducibility already, it is still too strict for many purposes of science. If we assign the psychological approach of behaviourism as one based on $\mathcal{C}2$-reducibility (see the following section), then we can say that Carnap (1956, 70) describes the problem as follows:

> "[T]he psychological movement of Behaviorism had, on the one hand, a very healthful influence because of its emphasis on the observation of behavior as an intersubjective and reliable basis for psychological investigations, while, on the other hand, it imposed too narrow restrictions."

Rather, according to Carnap (1956, 71), we need an even further weakening of the reductionist constraints:

> "In contrast to [the Behaviourist interpretation of mental concepts as being identified with behavioural patterns/dispositions], the interpretation of a psychological concept as a theoretical concept [...] does not identify the concept (the state or trait) with the pure disposition[.] The decisive difference is this: on the basis of the theoretical interpretation, the result of this or of any other test or, generally, of any observations, external or internal, is not regarded as absolutely conclusive evidence for the state in question; it is accepted only as probabilistic evidence, hence at best as a reliable indicator, i.e., one yielding a high probability for the state."

In particular, whereas it is characteristic of dispositional terms and bilateral reduction that they provide "conclusive evidence" for or against attributing a dispositional term (cf. Hempel 1958), theoretical concepts do not have this feature but have "at best evidence yielding a high probability" (cf. Carnap 1956, 69). In the context of the mental/physical distinction above, Carnap suggested to understand the 'reduction' of mental predicates as establishing them in the role of theoretical terms. We do so by connecting them to the empirical, observational, or physical by so-called "rules of correspondence" (cf. Carnap 1956, 39). What is the role of theoretical terms? Carnap says that it is to leave their empirical interpretation fruitfully open, i.e. to consider statements with such expressions as only providing an implicit and partial "definition" of them. In which sense can their incompleteness be fruitful? Carnap suggests that

> "we see at present the beginnings of similar developments [as in physics] in other fields of science, and there can be no doubt that here too the more comprehensive use of this method [of introducing terms by postulates without a complete interpretation] will lead in time to theories much more powerful for explanation and prediction than those theories which keep close to observables. Also in psychology, in these last decades, more and more concepts were used which show the essential features of theoretical concepts."

So, the third notion of reduction suggested by Carnap is this:

Definition *(reduction by establishing a theoretical concept) A mental predicate M is $\mathcal{C}3$-reducible to a set of physical predicates $\mathcal{P}$ iff M can be established as a theoretical term that can be connected to $\mathcal{P}$ by rules of correspondence that allow for powerful explanations and predictions.*

One way to spell out this implicit characterisation of theoretical terms (of $\mathcal{T}_{\mathcal{M}}$) by the help of rules of correspondence is that of the Ramsification of statements or theories with these terms (cf. Carnap 1966, 270). In Ramsifying such a statement, we replace all mental predicates (of $\mathcal{T}_{\mathcal{M}}$) by new higher order variables and existentially quantify over them. So, e.g., if we take the schema of a bilateral reduction sentence from above as a rule of correspondence between M and P_1, P_2, then a Ramsification of it would be, e.g.:

$$\exists X(P_1(x,t) \rightarrow (X(x) \leftrightarrow P_2(x,t))$$

This expression has no longer a mental or theoretical term; the higher order variable X is a placeholder for a structural node, marking the role of the mental or theoretical term, namely to link P_1 and P_2 in a particular way. As we have seen above, bilateral reduction sentences are creative and only partially eliminable, that is why we cannot provide an equivalent explicit definition of the involved mental or theoretical term. So, the term is not completely but at best partially (or implicitly or incompletely) defined. It is this incompleteness that makes it a possibly fruitful candidate for powerful explanations and predictions. Whether such structural nodes that link the empirical, observational, or physical predicates exist or not is, according to Carnap, evidentially not conclusive. However, we try to confirm it (cf. the above quoted aim of "yielding a high probability"). And, according to Carnap (1966, 270), if we take it to be true, we can also take the original claim containing the theoretical term to be true: "*[I]f* the Ramsey sentence is true, we must then understand the theoretical terms in such a way that the entire theory is true." Carnap (1966, 270) suggested to consider this implication between the Ramsification of a theoretical claim (^RTC) and the theoretical claim itself (TC), i.e.

$$^RTC \rightarrow TC$$

as the analytic content of the theoretical claim. The idea is that ^RTC has to be confirmed or undermined, but if it is confirmed, we can also expect that our respective theoretical claims are true. Whereas we call ^RTC the 'Ramsey sentence' of the theoretical claim TC, we call the implication $^RTC \rightarrow TC$ the 'Carnap sentence' of TC (cf. Lewis 1972). Note that the Carnap sentence plays an important role in an abductive inference: ^RTC contains next to logical only empirical terms; TC contains also theoretical terms. Given the Carnap sentence and the evidential confirmation of ^RTC, we can infer TC, so we infer a theory from empirical facts, which is characteristic of an abductive inference (cf. Schurz 2008; Feldbacher-Escamilla and Gebharter 2019).

As we have seen, bilateral reduction sentences are a particular form of rules of correspondence. The same holds for explicit definitions. In this sense, the constraint for $C3$ -reducibility is even weaker than those of $C2$- and $C1$-reducibility. However, that holds only with respect to the constraint of the logical form of such reductions (from explicit definition via bilateral reduction towards Ramsification). Contentwise, it is, furthermore, important that $C3$-reducibility aims for evidential *in*conclusiveness, because it is this feature that makes theoretical terms flexible enough to serve as vehicles for making (good inductive) generalisations (cf. Carnap 1956, 72f) and so allow for powerful explanations and predictions. For statements with explicitly definable terms, evidence is generally conclusive and for such statements with bilaterally reducible terms, evidence is conclusive for all test condition applications. In this sense, $C3$-reducibility and $C1$- as well as $C2$-reducibility can fall apart (and $C1$- as well as $C2$-reducibility might be no special case of $C3$-reducibility), simply because we can always formally transform explicit definitions to equivalent bilateral reduction sentences to—given the respective Carnap-sentence—equivalent Ramsifications; however, whereas the former are evidentially conclusive, the latter are not. And this holds also if we could provide for a particular Ramsification or bilateral reduction sentence an equivalent explicit definition (cf. Hempel 1958, he provides examples of $C1$-reducible terms that are still theoretical, because evidence for or against their application is inconclusive, 59f).

Now, as said, this *evidential inconclusiveness* in order to allow for powerful explanations and predictions is the key feature and main difference of reduction via establishing the role of a mental predicate (of $\mathcal{T}_{\mathcal{M}}$) as one of a theoretical term vs. reduction via a bilateral reduction sentence. In his *Two Dogmas*, Quine (1951, 38) also ascribes this feature of evidential (inconclusive) impact to Carnap and highlights its probabilistic nature:

> "Reductionism in its radical form [i.e. $\mathcal{C}1$-reduction] has long since ceased to figure in Carnap's philosophy. But the dogma of reductionism has, in a subtler and more tenuous form, continued to influence the thought of empiricists. The notion lingers that to each statement, or each synthetic statement, there is associated a unique range of possible sensory events such that the occurrence of any of them would add to the likelihood of truth of the statement, and that there is associated also another unique range of possible sensory events whose occurrence would detract from that likelihood."

As we have said, the understanding is that if we take any statement with a, e.g., mental term, then a $\mathcal{C}3$-reduction claims that we can establish the relevant role of it as that of a theoretical term. And one important feature of theoretical terms is their evidential inconclusiveness, i.e. that evidence has some probabilistic or confirmatory impact on the claim with the mental term (without fixing its truth or falsity). Although Carnap spelled out his approach of theoretical terms and their distinctions from "pure dispositions" later (cf. the overview in (Leitgeb and André 2020, Supplement E): 1956, 1958/1975, 1959/2000, 1966), already in his (1936) he hinted at this distinction (our emphasis):

- p.468: "Thesis of *Physicalistic Confirmability*: "Every descriptive *predicate* of the language of science is *confirmable* on the basis of observable thing-predicates."
- p.457: "[Definition:] A *predicate* 'P' is called *confirmable* (or completely confirmable, or *incompletely* confirmable) if 'P' is reducible (or completely reducible, or incompletely *reducible*, respectively) to a class of observable predicates."
- p.420: "We call [a *sentence*] *confirmable* if we know under what conditions the sentence would be confirmed."
- p.456: "[Definition:] A *sentence* S is called *confirmable* (or completely confirmable, or incompletely confirmable) if the confirmation of S is *reducible* (or completely reducible, or *incompletely* reducible, respectively) to that of a class of observable *predicates*."

We will focus here on this feature of evidential (inconclusive) confirmability of statements with theoretical terms, so we make it explicit as a meaning postulate amending the definition above:

Meaning Postulate (confirmatory amendment to $\mathcal{C}3$-reduction) If a mental predicate M is $\mathcal{C}3$-reducible to a set of physical predicates $\mathcal{P}$, then M-statements can be confirmed or undermined by evidence stated in terms of $\mathcal{P}$ only.

Focusing on this feature of reducibility by establishing the methodological role of a theoretical term, we can say that a positive feature of $\mathcal{C}3$-reductions is that they oftentimes might be easier to achieve (than $\mathcal{C}2$- and $\mathcal{C}1$-reductions), because of the practice of psychology and its confirmatory talk. On the other hand, this comes at the cost of a confirmatory methodology that is ultimately hard to articulate. However, as we will see in the next section, for our purposes in linking to the debate in the philosophy of mind, already

reference to the most general characteristics of the confirmatory methodology suffices in order to gain a relevant connection between philosophy of mind models and Carnapian reductionism.

To sum up, Carnap's reductionism comes in three stages:

- 1928: $\mathcal{C}1$-reducibility in terms of explicit definability (mainly in the *Aufbau*)
- 1936/37: $\mathcal{C}2$-reducibility in terms of bilateral reducibility (mainly in *Testability and Meaning*)
- 1950s and 60s: $\mathcal{C}3$-reducibility in terms of establishing the methodological role of theoretical concepts and the feature of (incomplete) confirmability (mainly in *The Methodological Character of Theoretical Concepts*)

We can also observe that this development expresses a weakening of reductionistic constraints, particularly if we focus on the formal constraints. Carnap himself described this development of his reductionistic account in the second preface (1961) to the *Aufbau* as follows:

> "In the sequel I want to indicate in what respects I have changed my position since I wrote the *Aufbau*. [...]
> One of the most important changes is the realisation that the reduction of higher level concepts to lower level ones cannot always take the form of explicit definitions [i.e. our $\mathcal{C}1$-reduction]; generally more liberal forms of concept introduction must be used. [...] These changes have been explained in [1936; 1937]. In that article I suggested the so-called reduction sentences as a more liberal form for the introduction of concepts [i.e. our $\mathcal{C}2$-reduction], which is especially suitable for dispositional concepts. Later on I considered [...] the introduction of 'theoretical concepts' through theoretical postulates and correspondence rules, and investigated the logical and methodological character of these concepts (cf. [1956]) [i.e. our $\mathcal{C}3$-reduction]. The correspondence rules connect the theoretical terms with observation terms. Thus the theoretical terms are interpreted, but this interpretation is always incomplete." (cf. Carnap 1928/2003, pp.viiif)

The quite general reconstruction of von Kutschera (1991) differs from ours inasmuch as he distinguishes between our $\mathcal{C}1$-reduction, lumps together $\mathcal{C}2$- and $\mathcal{C}3$-reduction, and speaks of another form of reduction to be found in Carnap (1963), namely the weakening of the constraint of intersubjective observability to subjective observability (while still upholding intersubjective confirmability). In fact, Carnap (1963, 882) claims that:

> Predicates designating properties that are only subjectively observable, though intersubjectively confirmable (as e.g., "angry", "having a toothache") may be introduced derivatively.

However, this weakening concerns a feature we did not make explicit in our reconstruction of $\mathcal{C}3$-reductions (although we could easily do so by amending the talk of "a set of physical predicates $\mathcal{P}$" in the definition of $\mathcal{C}3$-reduction above by "intersubjectively observable physical predicates" and then distinguish further a $\mathcal{C}4$-reduction by modifying "intersubjectively observable" to "subjectively observable"). von Kutschera (1991, 310) argues that this distinction collapses for Carnap, because for him, everything that can be observed subjectively can also be observed intersubjectively (e.g. via the observation of behaviour or via communication). For this reason, according to von Kutschera (1991, 311), it also makes no difference for different accounts of the philosophy of

mind. And since it plays no role there, and also not in our linking of Carnap's reductionism to the philosophy of mind debate in the next section, we do not explicitly introduce this further distinction here.

Let us provide a short illustration in order to better memorise this development by the help of a toy example, namely a "reduction" of the notion of *aggression*:

- $\mathcal{C}1$: *explicit definability*: x is aggressive at t *iff* x's serotonin level at t significantly exceeds the characteristic level for her/his type.

$$M(x,t) \quad \leftrightarrow \quad P(x,t)$$

- $\mathcal{C}2$: *bilateral reducability*: If x is *tested* by P_T at t, then x is aggressive iff x *reacts* the way P_R at t.

$$P_T(x,t) \;\rightarrow\; (M(x) \leftrightarrow P_R(x,t))$$

- $\mathcal{C}3$: *empirical confirmability* (focussing on one feature of $\mathcal{C}3$-reductions): there is empirical evidence P *confirming* that x is aggressive.

$$P(x,t) \text{ increases the probability of } M(x,t)$$

There is another important feature underlying all accounts of Carnap's reductionism from the very beginning: Carnap's distinction between internal and external questions. The main idea is described already in his (1928) and explained in more detail in his (1950; for a historical investigation cf. Verhaegh (2017)): scientists work with a linguistic framework; questions that can be answered within such a linguistic framework are internal questions; questions that cannot be answered within such a linguistic framework are external questions; since we can answer questions justifyably only with reference to a linguistic framework, external questions do not have a justifyable answer and, hence, are "pseudo-questions" and "pseudo-problems". For Carnap, this distinction is particularly relevant when it comes to questions of existence. If we, e.g., ask whether there is (i.e. exists) "a prime number greater than a hundred" (1950, sect.2), we can use the linguistic framework of mathematics to show that, e.g., from the axioms of mathematics we can deduce that there is a prime number greater than a hundred, e.g. 101. If we, however, ask whether there exist numbers, prime numbers, or even a prime number greater than a hundred *per se* (i.e. without reference to a linguistic framework; in philosophical talk we often use "really exist" for this), then we are asking a question that should be answered without reference to a linguistic framework, i.e. an external question, i.e. we are putting forward a pseudo-question. As we will see, with respect to the philosophy of mind, this distinction is relevant, because Carnap's accounts of reduction are always about reduction within a linguistic framework (to provide an explicit definition, a bilateral reduction sentence, or a rule of correspondence within a linguistic framework) and by this differ from what has been described in the previous section as ontological reduction. von Kutschera (1991, 306) described this underlying feature of Carnap's accounts of reduction as a "linguistic-pragmatic turn".

As we have spelt out all the necessary details and features of Carnap's accounts of reduction, we are now going to link his accounts to that of the philosophy of mind as outlined in the previous section.

4 Linking Carnap to the Philosophy of Mind

In section 2, we have stated that the predominant notion of 'reduction' in the philosophy of mind is Nagel's account of derivability by the help of coordinating definitions. We have also seen that, although he suggested different types of coordinating definitions, all of them have coextensionality and by this explicit definability in common. So, it is clear that this notion of 'reduction' coincides extensionally with Carnap's strongest account of $\mathcal{C}1$ -reduction (a difference remains, of course, with respect to the modal force). So, we can observe that in our big picture of the philosophy of mind landscape, reductivist accounts are actually also $\mathcal{C}1$-reductivist accounts. And, which should be particularly highlighted, non-reductivist accounts are actually non-$\mathcal{C}1$-reductivist accounts.

Now, let us see which accounts/models of philosophy of mind fit to Carnap's different accounts of reduction. Some of these mappings are pretty obvious, because either Carnap himself referred to them (as, e.g., with respect to behaviourism), they refer to Carnap (as, e.g., Herbert Feigl does) or there is an obvious overlap of the technical apparatus in use (as is, e.g., the case in Lewis's functionalism). von Kutschera (1991) has already discussed some relevant connections. He did so particularly with respect to type identity theory and $\mathcal{C}1$-reduction, and the realisation account of functionalism and $\mathcal{C}2$-reduction. However, we will argue in the subsequent section that his interpretation regarding $\mathcal{C}1$-reduction has to be corrected. We also think we can add to von Kutschera's analysis the discussion of supervenience physicalism (general supervenience) and $\mathcal{C}3$-reduction.

4.1 $\mathcal{C}1$-Reduction and Type Identity Theory

Let us have a quick view on von Kutschera's linking of type identity theory to $\mathcal{C}1$-reduction: he focuses particularly on Feigl's type identity account (cf. von Kutschera 1991, sect.2(1)). von Kutschera's reconstruction of Feigl's argument for the type identity of the mental and the physical is, roughly speaking, as follows: mental types and states are identical to physical/neurological types and states, simply because in a successful reductive enterprise we can define the former on the basis of the latter. In many cases it is common practice to conclude such an identity statement from such a definability statement. Hence, also in the case of psycho-physical definability we can conclude psycho-physical identity. To give an example, we know that the predicate 'is a bachelor' can be defined on the basis of 'is unmarried' and 'is a man' (this is a definability claim). From this we unproblematically conclude that bachelors are nothing else than unmarried men (this is an identity claim). Similarly we can conclude from the claim that a predicate for a mental state type can be defined with the help of predicates for neurophysiological state types (definability claim) that the mental state type is nothing else than a neurophysiological state type (identity claim). von Kutschera correctly points out that, regarding the co-extensionality/definability claim, such a reduction provides a philosophy of mind model for reduction in the sense of explicit definability as we find it in Carnap's *Aufbau* (our $\mathcal{C}1$-reduction). However, he then goes on to contrast Feigl's "metaphysical" account with Carnap's "linguistic-pragmatic" prerequisite, claiming that Feigl's metaphysical conclusion is not licensed because the definability claim in question is not analytic (as is the case for the bachelor example). The problem, according to von Kutschera, is that Feigl intends to infer an external (pseudo-) statement about the mental being identical with the neurophysiological. However, he cannot do so because the underlying definability statement is internal and only applies to a

mental-neurophysiological theory. So, what Carnap's $\mathcal{C}1$-reductivist account with the linguistic-pragmatic restriction to internal questions and answers allows would be at most an internal identity claim that *according to a mental-neurophysiological theory/framework* the mental is identical to the neurophysiological (compare this to our discussion in the previous section regarding the existence of a prime number greater than a hundred *according to mathematics* vs. the existence of prime numbers *per se*).

We fully agree with von Kutschera's Carnapian analysis of his reconstruction of Feigl's reasoning. However, we think that this reasoning can hardly be attributed to Feigl but rather makes up for a straw man version of Feigl's account. The reason is simply that, true, Feigl speaks of identification occasionally very generally and without specifying a particularly framework. This is the case, e.g., in the following passage:

> "The identity thesis which I wish to clarify and to defend asserts that the states of direct experience which conscious human beings 'live through,' and those which we confidently ascribe to some of the higher animals, are identical with certain (presumably configurational) aspects of the neural processes in those organisms." (cf. Feigl 1958b, 446)

However, in general Feigl takes very much care of contextualising his identity statement (which he often labels as 'ψ-ϕ identification'). So, e.g., in (Feigl 1958b, sect.V(D)) he takes "the terminological liberty of speaking of different kinds of identity, viz., (1) logical, (2) empirical; and under (2) I shall distinguish (a) accidental, (b) nomological, (c) theoretical identities" (p.440). Afterwards he makes clear that whenever he speaks of $\psi - \phi$ identification, he understands it as an empirical (nomological or theoretical) identity (e.g. on p.442: "*if* this physicalistic program can be carried out, then there would be something like an *empirical* identification"; on p.445: "I shall now present, as explicitly as I can, the reasons for an empirical identification of raw feels with neural processes"; on p.448 he speaks of the "the empirical character of the identification", etc.). So, it seems to be clear that what Feigl had in mind was no external or ontological (in the sense of what "really" exists) identification but only an internal identification on the basis of a psycho-neurophysiological framework. And such a relativised identity claim is, in fact, licensed by the relativised definability claim: Given a psycho-neurophysiological framework (Feigl speaks of "the neurophysiology of the future (3000 A.D.?)" in his 1958b, 442), terms and predicates for mental states and state types can be defined on the basis of terms and predicates for neurophysiological states and state types and, hence, they can be also identified given such a framework (not *per se*). In this sense we think that type identity theory of, e.g., Feiglean provenience serves as a well fitting philosophy of mind model for $\mathcal{C}1$-reduction even if one takes into account the "linguistic-pragmatic" restriction.

Before we come to the next version of reductionism, let us briefly see how Carnap's strongest form of reductionism in the sense of explicit definability in general copes with the conditions of adequacy for physicalism in the philosophy of mind as put forward in section 2: As for $\mathcal{A}_{\mathcal{A}}$ (autonomy), it is pretty obvious that Carnap's remarks on the "unity of science" and "the impossibility of an intersubjective language not included in the physical language" (cf. Carnap 1995, 66; German original in Carnap 1931, 447f) directly counters this condition. Regarding $\mathcal{A}_{\mathcal{C}}$ (causal exclusion), a complete reduction of mental predicates to physical predicates allows also to translate statements about mental causation into statements about physical causation; traditional problems of the mental-physical interaction are in the early Carnapian reductivist framework oftentimes analysed as pseudo-problems (cf., e.g., Carnap 1928/2003, sect.162 About Mind-Body Dualism). Regarding $\mathcal{A}_{\mathcal{K}}$ (kinds), it is important to note that talk of "natural kinds" became particularly important with respect to

inductive generalisation, something that is not so present in the early reductive period of Carnap and his reliance on *verificationism* (only starting with Carnap 1936, we find a shift from verification towards confirmation/induction; cf. particularly 425); and although he tried to provide a formal criterion for defining kinds (such as colour terms) in his *Aufbau* by characterising the method of *quasi-analysis* (cf. Carnap 1928/2003, sect.71), it seems that his highlighting of the importance of co-extensionality of the result of a rational reconstruction just marginalises the role of such kinds: whichever definiens generates a co-extensional outcome seems to be fine—be it a natural kind or not. Regarding $\mathcal{A}_{\mathcal{M}}$ (multiple), there is no indication in Carnap's early work that would rule out disjunctions from analysis, so, a disjunctive treatment of multiple realisation seems per default favourable. And, finally, again by stressing the central role of physics, perhaps not so much in the *Aufbau* but particularly in the course of the *Protokollsatzdebate* where mainly Otto Neurath convinced Carnap to switch to a physical basis (cf. 1931, 452, fn.1, for a reference of Carnap to Neurath), it seems that the aspiration of Carnap's early reductionism was to provide a complete analysis of the mental, so, if this form of reductionism succeeds, there should be no mental residue ($\mathcal{A}_{\mathcal{R}}$).

4.2 $\mathcal{C}2$-Reduction and Functionalism

When working out his notion of $\mathcal{C}2$-*reduction* in his (1936 and 1937), Carnap referred to the the behaviouristic branch of psychology (cf. p.454) as importantly stressing the role of observable behaviour. Although, as we have seen in our reconstruction of Carnap's motivation to move from $\mathcal{C}2$-reduction to $\mathcal{C}3$-reduction and his criticism of *Behaviourism* in section 3, Carnap later on assessed the behaviouristic account as providing a valuable incentive to shift a field more towards the stage of empirical confirmation, but that a too strict demand of sticking to the behaviouristic or more generally the operational methodology (by, e.g., even identifying the meanings of terms with methods of measurement or patterns of behaviour) as ultimately hindering scientific progress, in the context of $\mathcal{C}2$-reductionism he worked out formal details of behaviourism by providing the general schema of a bilateral reduction sentence.

In the philosophy of mind, respective positions were established only later. One important branch that maps quite well to $\mathcal{C}2$-reduction is that of so-called "analytical or logical behaviourism", which states that mental terms and predicates can be reconstructed as behavioural dispositions or tendencies. E.g. the account of Gilbert Ryle and his in *The Concept of Mind* might be mentioned here as one of providing an ordinary language behaviouristic analysis. For Ryle, we should think about mental terms and predicates as we think about dispositions like *being able to speak French*, where "we expect no more than that [someone ...] copes pretty well with the majority of ordinary French-using and French-following tasks" (cf. 1949, 124)

Behaviourism of this sort is sometimes also described as *weak (or mono-causal) functionalism*, because it aims at the individual characterisation of mental terms and predicates by the help of behavioural test-reaction pairs (cf. von Kutschera 1991, 314). More complex is *strong functionalism*, which tries to characterise mental terms and predicates by the help of the causal role of the respective states and types in a causal setup. The idea is not to focus on individual concepts, but rather on such concepts in a general (causal) network. The tendency is to establish the methodological role of such terms and predicates as one of theoretical concepts; in this sense, strong functionalism would fit better to $\mathcal{C}3$-reduction. However, as we will see in a bit, strong functionalism—particularly the version we

have in mind here—sometimes comes with a claim of unique attributability or "functional definability" (in the sense that the concepts are characterised by their role in a general network of concepts, and each concept has such unique features that it can be individually functionally defined). This treatment as one of $\mathcal{C}2$-reduction is also suggested by several analogies we find between the dispositional and the functional approach: like *dispositional concepts* introduced by bilateral reduction sentences stress the role of dispositions (tendencies, capacities etc.) to link tests and reactions, functionalism stresses the *functional input–output role* of the mental; like dispositional statements have an empirical as well as a theoretical component (empirical are the test and reaction conditions; theoretical is the dispositional linking), so do mental terms according to functionalism (empirical are input and output; theoretical is the function); also, what both have in common is that they stress that the same phenomenon can be *realised* by different physical setups; the dispositional analysis does so by the help of multiple test-reaction pairs; likewise, functionalism stresses the possibility of different forms of realisation such as, e.g., the possibility to differently realise pain as C-fibers firing in humans, as opening of D-valves in Martians, as something else in octopi etc. Let us provide a little bit more details on this by reference to the account of Lewis (1972). Lewis argues for the account that "a mental state *M* (say, an experience) is definable as the occupant of a certain causal role *R*–that is, as the state, of whatever sort, that is causally connected in specified ways to sensory stimuli, motor responses, and other mental states." (p.250). The main idea is that we take a complete causal description that contains mental vocabulary such as, e.g., a folk/psychological account of frustration and aggression: "If *FRUSTRATION*(x), then *AGGRESSION*(x), where, e.g., *lock-out*(x) leads to *FRUSTRATION*(x), and, e.g., usually *AGGRESSION*(x) shows via *shout*(x)" (mental expressions being upper case; observational expressions being lower case). Now, as outlined in section 3, we Ramsify (without existential quantifiers) the statement and end up with

$$(X_1(x) \rightarrow X_2(x)),\ \&\ (lock\ -\ out(x) \rightarrow X_1(x))\ \&\ (X_2(x) \rightarrow shout(x))$$

so, "we get a formula in which only [observational]-terms appear [... and] any n-tuple of entities which satisfies this formula is a realization of the theory" (Lewis 1972, 253). Now, Lewis (1972, 254) goes on to suggest to aim for a "functional definition" by not only asking for a Ramsification and Carnap sentence analysis of the meaning of mental terms, but by putting forward a unique-Ramsification for each mental term (cf. the technical deviation in von Kutschera 1991, 314; we use $\exists!$ for "there exists exactly one"):

$$\exists! X_1 \exists X_2 (X_1(x) \rightarrow X_2(x))\ \&\ (\text{lock - out}(x) \rightarrow X_1(x))\ \&\ (\text{X}_2(\text{x}) \rightarrow \text{shout(x)})$$

$$\exists! X_2 \exists X_1 (X_1(x) \rightarrow X_2(x)) \& (lock\text{ - }out(x) \rightarrow X_1(x)) \& (\text{X}_2(x) \rightarrow \text{shout}(x))$$

If we take the unique-versions of Carnap sentences on the basis of these statements, we end up with a definition of *FRUSTRATION* and *AGGRESSION* on the basis of observational terms only (*FRUSTRATION* and *AGGRESSION* marking the respective nodes in the causal network). In this sense, also strong functionalism provides a philosophy of mind model for $\mathcal{C}2$-reduction (in the strongest form with reduction sentences that would even allow for a $\mathcal{C}1$ -reduction).

Let us also briefly think about the profile of Carnap's $\mathcal{C}2$-reductionist account regarding the conditions of adequacy and whether it is similar to that of functionalism: As for $\mathcal{A}_{\mathcal{A}}$ (autonomy), since mental terms cannot be completely but only partially eliminated by the help of bilateral reduction sentences, there remains some autonomy of the higher level

science, namely with respect to all cases where no suitable test condition applies. Regarding $\mathcal{A}_{\mathcal{C}}$ (causal exclusion), Carnap is pretty much quiet in the sense that he denies the scientific relevance of causal statement. So, e.g., in (Carnap 1928/2003, sect.165; and Carnap 1936, 471) he assigns causal talk to the "material idiom" which needs to be paraphrased in talking about bilateral reduction itself. Since Lewis is also close to Carnap's $\mathcal{C}3$-reductionism, one might consider this causal analysis also as one that might have been accepted by Carnap. Regarding $\mathcal{A}_{\mathcal{K}}$ (kinds), again, as we have discussed in subsection 4.1, in this phase of Carnap's development natural kinds did not play an important role, for which reason he did not say anything about it; however, due to the technically similar treatment, it seems reasonable to apply the approach of Lewis for a Carnapian account of natural kinds, which, as we outlined in section 2, arguably is able to deal with this condition of adequacy. Regarding $\mathcal{A}_{\mathcal{M}}$ (multiple), Carnap's talk about empirical creativity also via squaring test-reaction pairs makes clear that his account is meant to allow multiple realisations/manifestations, i.e. that a disposition can be attributed on the basis of several test-reaction pairs. Finally, coming to $\mathcal{A}_{\mathcal{R}}$ (residue), we saw that partial eliminability was a feature with respect to autonomy, however with respect to the reduction of mental terms and predicates it leaves open a gap/residue, namely the non-reducibility with respect to all cases where no suitable test-reaction conditions can be found.

4.3 $\mathcal{C}3$-Reduction and Supervenience Physicalism

Let us now come to empirical confirmability as one of the main desiderata of $\mathcal{C}3$-reducibility in the sense of establishing the methodological role of mental terms as theoretical terms that can, finally, serve as vehicles of good inductive generalisations. Now, which approach of the philosophy of mind could serve as a model of this form of reduction? Lewis (1972) and $\mathcal{C}3$-reductions as proposed in (Carnap 1956; 1966) are based on the same technical apparatus: Ramsification, and identification of meaning with the Carnap sentence. We have seen in subsection 4.2 that Lewis (1972) also suggests a stronger form of interpreting mental vocabulary via unique existential quantification in the Ramsey and Carnap sentences or, what he calls, *functional definability*. If we take as a weaker version ordinary Ramsification and meaning identification via the Carnap sentence, we end up with a model of $\mathcal{C}3$-reduction (Lewis (1972) speaks also of *theoretical terms* but does not highlight their confirmatory role).

Here we want to suggest, however, a further connection, namely one between supervenience physicalism and empirical confirmability. As we have seen in section 2, supervenience physicalism only postulates general supervenience (that might be spelled out differently and amended with an explanation of why supervenience holds by such accounts as, e.g., realisation and grounding). Now, there exists a multitude of notions of *supervenience* (cf. McLaughlin 1995). We want to start with a weak and purely extensional version (cf., e.g. von Kutschera 1991, 316): Property supervenience claims that an M property supervenes on a P property if all objects that are alike with respect to the P property are also alike with respect to the M property:

Definition *(property supervenience) A property M supervenes on a property P iff it holds for all objects x and y*:

$$(P(x) \leftrightarrow P(y)) \rightarrow (M(x) \leftrightarrow M(y))$$

Most of the time we face not only questions of supervenience with respect to single properties, but with respect to sets of properties. We can generalise the notion of *property supervenience* to a notion of the *supervenience of properties* via an indistinguishability with respect to the sets of properties:

Definition *(supervenience of properties) A set of properties $\mathcal{M}$ supervenes on a set of properties $\mathcal{P}$ iff it holds for all objects x and y: If it holds for all $P \in \mathcal{P}$: $P(x) \leftrightarrow P(y)$, then it holds for all $M \in \mathcal{M}$: $M(x) \leftrightarrow M(y)$.*

If we define identity with respect to a set of properties $\mathcal{F}$, i.e. $=_{\mathcal{F}}$, as:

$$x =_{\mathcal{F}} y \text{ iff } \forall F \in \mathcal{F} : \; F(x) \leftrightarrow F(y)$$

Then we can express supervenience of properties as follows:

$$\forall x \forall y : \; (x =_{\mathcal{P}} y \rightarrow x =_m y)$$

I.e.: $\mathcal{P}$-indistinguishability implies $\mathcal{M}$-indistinguishability, or, via contraposition: $\mathcal{M}$-distinguishability implies $\mathcal{P}$-distinguishability:

$$\forall x \forall y : \; (x \neq_m y \rightarrow x \neq_{\mathcal{P}} y)$$

Trivial cases of such extensional supervenience are, e.g., co-extensional properties that supervene on each other but also coarse-grained properties (like being in the equivalence class of the *Urmeter* up to a difference of $1cm$) that supervene on more fine-grained properties (like being in the equivalence class of the *Urmeter* up to a difference of $1mm$). We can also gain intensional versions of supervenience relations (with modal force) by adding a modal operator $\Box$:

$$\Box \forall x \forall y : \; (x =_{\mathcal{P}} y \rightarrow x =_m y)$$

Typically, supervenience is supposed to be a metaphysical relation, that is why the $\Box$ typically stands for some "realistic" or "worldly" modality such as *nomological necessity* or *metaphysical necessity*. However, in principle it can also stand for a *conceptual* or *logical* (analytical) modality. For our interpretation of $\mathcal{C}3$-reduction we want to suggest an epistemic/probabilistic reading, however, we do so not by defining a particular epistemic modality (so we stick to the extensional version of the *supervenience of properties* as stated above), but by restricting the property sets $\mathcal{P}$ and $\mathcal{M}$ to epistemic/probabilistic properties, namely that of *probabilistic distinction/update*. In particular we have in mind the claim that there is no probabilistic distinction/update on the side of the mental (a statement $S_{\mathcal{M}}$ in terms of mental predicates $\mathcal{M}$), if there is no probabilistic distinction/update on the side of the physical (a statement $S_{\mathcal{P}}$ in terms of physical predicates $\mathcal{P}$):

$$Pr_i(S_M) \neq Pr_o(S_M) \quad \rightarrow \quad Pr_i(S_{\mathcal{P}}) \neq Pr_o(S_{\mathcal{P}})$$

If we take Pr_i to be a prior probability distribution and Pr_o to be a posterior probability distribution, then there is also a quite natural interpretation of this dependence of the mental on the physical, namely in terms of Bayesian updating on (physical) evidence:

$$\underbrace{Pr_i(S_m)}_{Pr_i(S_m)} \neq \underbrace{Pr_o(S_M)}_{Pr_o(S_m)=Pr_i(S_m|S_{\mathcal{P}})} \rightarrow \underbrace{Pr_i(S_{\mathcal{P}})}_{Pr_{prior}(S_{\mathcal{P}})} \neq \underbrace{Pr_o(S_{\mathcal{P}})}_{Pr_o(S_{\mathcal{P}})=Pr_i(S_{\mathcal{P}}|S_{\mathcal{P}})=1}$$

Which is *Bayesian updating* on $S_{\mathcal{M}}$ upon *learning* $S_{\mathcal{P}}$. In this sense, general *supervenience* as presupposed by *supervenience physicalism* can be understood as a general model for a $\mathcal{C}3$-reductionistic constraint of confirmation: no confirmation/undermination of the mental/theoretical, if there is no confirmation/undermination of the physical/empirical.

Let us conclude this section with a short reflection on how Carnap's account of empirical confirmability fares with respect to the conditions of adequacy for theorising in the philosophy of mind. As for $\mathcal{A}_{\mathcal{A}}$ (autonomy), Carnap is very explicit in his later writings about the autonomous status of sciences such as biology, sociology, psychology etc. Although Carnap (1956) still outlines a *physicalistic future* where "micro-physiology may be based on micro-physics" (cf. p.74), his tone is already very tamed ("may") and even if such a phase will be achieved (we have seen above that Feigl speculated about 3000 A.D.), there will be a very long period of autonomous psychology that establishes its own methodological concepts and undergoes "analogous [fruitful] developments" as was the case with physics (cf. p.74). Also regarding $\mathcal{A}_{\mathcal{C}}$ (causal exclusion), Carnap underwent quite some changes from the early syntactical phase, via the "rehabilitation" of *semantics* and subsequently the modal approach acceptable even for empiricists. Carnap himself worked on modality (1947) and in his (1956) he states openness to the use of "causal modalities" even if their inclusion would require a considerably more complicated set of rules of logical deduction (cf. p.42). Of course he did not delve more into this with respect to mental causation. However, since, as we have argued here, supervenience is also an underlying model of empirical confirmability, the account of $\mathcal{C}3$-reduction probably inherits the supervenience-based problem of mental causation. Regarding $\mathcal{A}_{\mathcal{K}}$ (kinds), we just need to refer to the role of theoretical terms as outlined in Carnap (1956), namely to allow for good inductive generalisations in order to provide successful explanations and predictions, a main feature of natural kinds. Regarding $\mathcal{A}_{\mathcal{M}}$ (multiple), since theoretical concepts are even more loosely linked to empirical terms than dispositional terms are via bilateral reduction, and the latter allow for multiple realisation, also states or structures expressed by theoretical terms should allow for multiple realisation; we have also seen that the account of Lewis (1972) which is structurally very similar to $\mathcal{C}3$-reduction allows for multiple realisation. Finally, with respect to $\mathcal{A}_{\mathcal{R}}$ (residue), we can observe that, since there is no longer an aim of complete subordination of one domain to the other, but a parallel development of establishing theoretical terms (in physics as in psychology), the question of mental residue is no longer an issue.

5 Conclusion

We have seen that Carnap's development regarding reductionism spans from the strong programme of explicit definability ($\mathcal{C}1$-reduction) via bilateral reducibility ($\mathcal{C}2$-reduction) towards empirical confirmability ($\mathcal{C}3$-reduction). We have argued that type identity theory provides a model for $\mathcal{C}1$-reduction, behaviourism and a strong version of functionalism provide a model for $\mathcal{C}2$-reduction, and supervenience physicalism provides a model for $\mathcal{C}3$ -reduction. We also gave independent reasons from Carnap's philosophy of science for the adequacy profiles as summarised in Table 2. As we see by comparison of Table 1 and Table 2, the profiles of the single $\mathcal{C}$-reductive accounts match those of the linked philosophy of mind models ($\mathcal{C}1$-reduction linked to a disjunctive version of type identity theory).

So much for the direction from the philosophy of mind to Carnap's accounts of reduction. However, we think that there is also a moral to be drawn for the other

Table 2 Overview of Carnap's different positions on reductionism and how they account for conditions of adequacy to be found within the philosophy of mind

Condition of adequacy		$\mathcal{C}1$-reduction	$\mathcal{C}2$-reduction	$\mathcal{C}3$-reduction
$\mathcal{A}_{\mathcal{A}}$	(autonomy)	−	+	+
$\mathcal{A}_{\mathcal{C}}$	(causal excl.)	+	+	−
$\mathcal{A}_{\mathcal{K}}$	(kind)	−	+	+
$\mathcal{A}_{\mathcal{M}}$	(multiple)	+	+	+
$\mathcal{A}_{\mathcal{R}}$	(residue)	+	−	+

direction, from Carnap's accounts to the philosophy of mind. To quickly outline the idea: If we find for several physicalist accounts of the philosophy of mind a fitting reductive Carnapian counterpart, one might question whether the distinction between reductive (type identity) and non-reductive (token identity) accounts is well-founded. Rather, this distinction is based only on the notion of Nagelean reduction but disregards other possible conceptions such as Carnapean $\mathcal{C}2$- or $\mathcal{C}3$-reduction. Since the latter can be interpreted as a gradual weakening of reductive constraints, it might be more fitting to frame the physicalist accounts as amounting more or less to gradually stronger or weaker notions of reduction (since we did not discuss non-physicalist accounts here, we stay neutral with respect to a revisionary view for this domain). The revised picture our investigation suggests is depicted in Fig. 2. Note that our suggestion is also backed by recent other investigations in this field as, e.g., by Hüttemann (2023) who also speaks of "'non-reductive physicalism' as a misnomer" due to its restriction to Nagelean reduction. Taken together, these insights suggest that the time has come to reconsider the concept of reduction and its role in debates within the philosophy of mind.

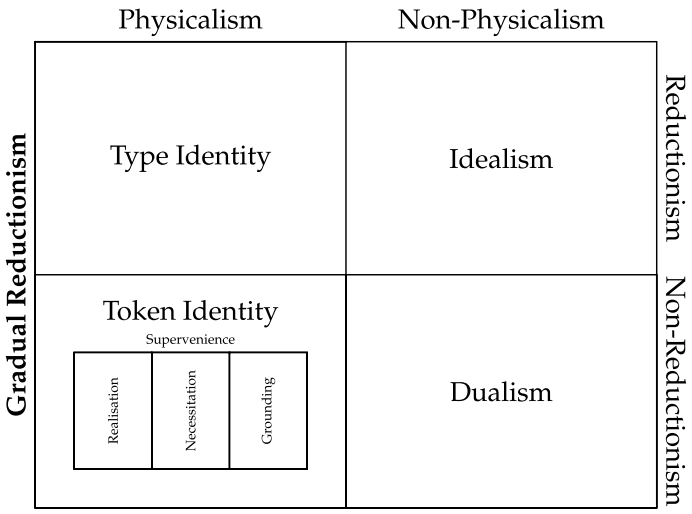

Fig. 2 Positions in the philosophy of mind: a revised picture; instead of distinguishing reductive from non-reductive physicalist accounts, we suggest to distinguish physicalist accounts gradually with respect to reduction

Acknowledgements For helpful discussions on this topic we would like to thank Alexander Gebharter, Vera Hoffmann-Kolss, Andreas Hüttemann, Jan Michel, Raphael van Riel, Gerhard Schurz and Corina Strössner.

Funding Open Access funding enabled and organized by Projekt DEAL. This research is funded by DFG, research unit FOR 2495, research grant SCHU 1566/11-2, as well as grant number 545054032 (Rudolf Carnap, the Problem of Induction, and the Choice of Scientific Frameworks).

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose. The authors have no Conflict of interest to declare that are relevant to the content of this article. All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript. The authors have no financial or proprietary interests in any material discussed in this article.

Ethical Approval The authors comply with the ethical standards of the journal.

Human or Animal Rights This article does not contain any studies with human participants or animals performed by any of the authors.

References

Carnap, Rudolf. 1928/1961. Der logische Aufbau der Welt. Scheinprobleme in der Philosophie. Hamburg: Felix Meiner.

Carnap, Rudolf. 1928/2003. The Logical Structure of the World and Pseudoproblems in Philosophy. Open Court Classics. Illinois: Open Court.

Carnap, Rudolf. 1932. "Psychologie in physikalischer Sprache". English. In: Erkenntnis 3:107–142. http://www.jstor.org/stable/20011672.

Carnap, Rudolf. 1934/2001. Logical Syntax of Language. London: Routledge.

Carnap, Rudolf. 1958/1975. "Observation Language and Theoretical Language". In: Rudolf Carnap, Logical Empiricist. Ed. by Hintikka, Jaakko. Dordrecht: Reidel Publishing Company.

Carnap, Rudolf. 1959/2000. "Theoretical Concepts in Science". Rudolf Carnap's 'Theoretical Concepts in Science'. Ed. by Psillos, Stathis. Studies in History and Philosophy of Science Part A. 31(1), 151–172. https://doi.org/10.1016/S0039-3681(99)00031-X.

Carnap, Rudolf. 1963. Replies and Systematic Exposition. In: *The Philosophy of Rudolf Carnap*, ed. Paul Schilpp, 858–1013. Arthur. La Salle: Open Court.

Carnap, Rudolf. 1928. *Scheinprobleme der Philosophie*. Berlin: Weltkreis.

Carnap, Rudolf. 1931. Die physikalische Sprache als Universalsprache der Wissenschaft. *Erkenntnis* 2: 432–465.

Carnap, Rudolf. 1936. Testability and Meaning. *Philosophy of Science* 3 (4): 419–471. https://doi.org/10.1086/286432.

Carnap, Rudolf. 1937. Testability and Meaning - Continued. *Philosophy of Science* 4 (1): 1–40. https://doi.org/10.1086/286443.

Carnap, Rudolf. 1947. *Meaning and Necessity: A Study in Semantics and Modal Logic*, 2nd ed. Chicago: University of Chicago Press.

Carnap, Rudolf. 1950. Empiricism, Semantics, and Ontology. *Revue Internationale de Philosophie* 4 (2): 20–40.

Carnap, Rudolf. 1956. The Methodological Character of Theoretical Concepts. In *Minnesota Studies in the Philosophy of Science*, vol. I, ed. Herbert Feigl. The Foundations of Science and the Concepts of Psychology and Psychoanalysis, 38–76. Minneapolis: University of Minnesota Press.
Carnap, Rudolf. 1959. Psychology in Physical Language. In *Logical positivism*, ed. Alfred J. Ayer, 165–198. Glencoe: Library of philosophical movements. The Free Press.
Carnap, Rudolf. 1966. Philosophical Foundations of Physics. In *An Introduction to the Philosophy of Science*, ed. Martin Gardner. New York: Basic Books.
Carnap, Rudolf. 1995. The Unity of Science. In *Translated with an Introduction by Max Black*, ed. Rudolf Carnap. Bristol: Thoemmes Press.
Clapp, Lenny. 2001. Disjunctive Properties. Multiple Realizations. *The Journal of Philosophy* 98 (3): 111–136.
Feigl, Herbert, ed. 1958a. Minnesota Studies in the Philosophy of Science. Volume II. Concepts, Theories, and the Mind-Body Problem. Minneapolis: University of Minnesota Press.
Feigl, Herbert, ed. 1958b. "The 'Mental' and the 'Physical' ". In: Minnesota Studies in the Philosophy of Science: Volume 02: Concepts, Theories, and the Mind-Body Problem, 370–497.
Feldbacher-Escamilla, Christian J., and Alexander Gebharter. 2019. Modeling Creative Abduction Bayesian Style. *European Journal for Philosophy of Science* 9 (1): 1–15. https://doi.org/10.1007/s13194-018-0234-4.
Fodor, Jerry A. 1974. Special Sciences (or: The disunity of science as a working hypothesis). *Synthese* 28 (2): 97–115. https://doi.org/10.1007/BF00485230.
Gebharter, Alexander. 2017. Causal Exclusion and Causal Bayes Nets. *Philosophy and Phenomenological Research* 95 (2): 353–375. https://doi.org/10.1111/phpr.12247.
Heidelberger, Michael. 2003. The Mind-Body Problem in the Origin of Logical Empiricism. In *Logical Empricism*, ed. Paolo Parrini, Wesley C. Salmon, and Merrilee H. Salmon, 233–262. Historical and Contemporary Perspectives: University of Pittsburgh Press, Pittsburgh.
Hempel, Carl G. 1958. The Theoretician's Dilemma. In *Minnesota Studies in the Philosophy of Science*, vol. II, ed. Herbert Feigl. Concepts, Theories, and the Mind-Body Problem, 37–98. Minneapolis: University of Minnesota Press.
Hoffmann-Kolss, Vera, and Nicole Rathgeb, eds. 2023. *Handbuch Philosophie des Geistes*. Stuttgart: J. B. Metzler.
Hüttemann, Andreas. 2023. Nichtreduktiver Physikalismus. In *Handbuch Philosophie des Geistes*, ed. Vera Hoffmann-Kolss and Nicole Rathgeb. Stuttgart: J. B. Metzler.
Kim, Jaegwon. 1989. Mechanism, Purpose, and Explanatory Exclusion. *Philosophical Perspectives* 3: 77–108.
Kim, Jaegwon. 2003. Logical Positivism and the Mind-Body Problem. In *Logical Empricism*, ed. Paolo Parrini, Wesley C. Salmon, and Merrilee H. Salmon. Historical and Contemporary Perspectives: University of Pittsburgh Press, Pittsburgh.
Kim, Jaegwon. 2005. *Physicalism, or Something Near Enough*. Princeton: Princeton University Press.
Kroedel, Thomas, and Moritz Schulz. 2016. Grounding Mental Causation. *Synthese* 193 (6): 1909–1923. https://doi.org/10.1007/s11229-015-0820-3.
Leitgeb, Hannes and Carus, André. 2020. "Rudolf Carnap". In: The Stanford Encyclopedia of Philosophy (Spring 2023 Edition). Ed. by Zalta, Edward N. Stanford University: Metaphysics Research Lab. https://plato.stanford.edu/archives/spr2020/entries/carnap/.
Levin, Janet. 2002. Is Conceptual Analysis Needed for the Reduction of Qualitative States? *Philosophy and Phenomenological Research* 64 (3): 571–591. https://doi.org/10.1111/j.1933-1592.2002.tb00161.x.
Levin, Janet. 2022. *The Metaphysics of Mind*. Cambridge: Cambridge University Press.
Lewis, David K. 1966. An Argument for the Identity Theory. *The Journal of Philosophy*. https://doi.org/10.2307/2024524.
Lewis, David K. 1972. Psychophysical and Theoretical Identifications. *Australasian Journal of Philosophy* 50 (3): 249–258. https://doi.org/10.1080/00048407212341301.
Lewis, David K. 1980. Mad Pain and Martian Pain. In *Readings in Philosophy of Psychology*, ed. Ned Block, 216–232. Cambridge: Harvard University Press.
Lewis, David K. 1994. Reduction of Mind. In *A Companion to Philosophy of Mind*, ed. Samuel Guttenplan, 412–431. Oxford: Blackwell Publishers.
List, Christian, and Peter Menzies. 2009. Nonreductive Physicalism and the Limits of the Exclusion Principle. *The Journal of Philosophy* 106 (9): 475–502. https://doi.org/10.5840/jphil2009106936.
Lycan, William G. 1987. *Consciousness*. Cambridge: MIT Press.
McLaughlin, Brian. 1995. Varieties of Supervenience. In *Supervenience: New Essays*, ed. Elias E. Savellos and Umit D. Yalçin, 16–59. Cambridge: Cambridge University Press.

McLaughlin, Brian. 2006. Is Role-Functionalism Committed to Epiphenomenalism? *Journal of Consciousness Studies* 13 (1–2): 39–66.
Michel, Jan G. 2023. Identitätstheorie und Eliminativismus. In: *Handbuch Philosophie des Geistes*, ed. Vera Hoffmann-Kolss and Nicole Rathgeb. Stuttgart: J. B. Metzler.
Moore, Dwayne. 2011. Role Functionalism and Epiphenomenalism. *Philosophia* 39 (3): 511–525. https://doi.org/10.1007/s11406-011-9302-0.
Nagel, Ernest. 1961. The Structure of Science. Problems in the Logic of Scientific Explanation. New York: Harcourt, Brace & World, INC.
O'Conaill, Donnchadh. 2018. Grounding, Physicalism and Necessity. *Inquiry* 61 (7): 713–730. https://doi.org/10.1080/0020174X.2017.1385524.
Pap, Arthur. 1958. Disposition Concepts and Extensional Logic. In *Minnesota Studies in the Philosophy of Science*, vol. II, ed. Herbert Feigl. Concepts, Theories, and the Mind-Body Problem, 196–224. Minneapolis: University of Minnesota Press.
Parrini, Paolo, Wesley C. Salmon, and Merrilee H. Salmon, eds. 2003. *Logical Empricism*. Pittsburgh: Historical and Contemporary Perspectives. University of Pittsburgh Press.
Place, U.T. 1956. Is Consciousness a Brain Process? *British Journal of Psychology* 47 (1): 44–50. https://doi.org/10.1111/j.2044-8295.1956.tb00560.x.
Polger, Thomas W. 2011. Are Sensations Still Brain Processes? *Philosophical Psychology* 24 (1): 1–21. https://doi.org/10.1080/09515089.2010.533263.
Putnam, Hilary. 1960. Minds and Machines. In *Dimensions of Minds*, ed. Sidney Hook, 138–164. New York: New York University Press.
Putnam, Hilary. 1967. Psychological Predicates. In *Art, Mind, and Religion*, ed. W.H. Capitan and D.D. Merrill, 37–48. Pittsburgh: University of Pittsburgh Press.
Quine, Willard van Orman. 1951. Two Dogmas of Empiricism. *The Philosophical Review* 60 (1): 20–43.
Ryle, Gilbert. 1949. *The Concept of Mind*. Chicago: University of Chicago Press.
Schneider, Susan. 2012. Why Property Dualists Must Reject Substance Physicalism. *Philosophical Studies* 157 (1): 61–76. https://doi.org/10.1007/s11098-010-9618-9.
Schurz, Gerhard. 2008. Patterns of Abduction. *English. In: Synthese* 164 (2): 201–234. https://doi.org/10.1007/s11229-007-9223-4.
Schurz, Gerhard. 2014. Criteria of Theoreticity: Bridging Statement and Non-Statement View. *Erkenntnis* 79 (8): 1521–1545. https://doi.org/10.1007/s10670-013-9581-x.
Shoemaker, Sydney. 1975. Functionalism and Qualia. *Philosophical Studies* 27 (5): 291–315. https://doi.org/10.1007/bf01225748.
Smart, J.J.C. 1959. Sensations and Brain Processes. *The Philosophical Review* 68 (2): 141–156. https://doi.org/10.2307/2182164.
Stoljar, Daniel. 2010. *Physicalism*. London: Routledge.
van Riel, Raphael, and Robert Gulick. 2019. Scientific Reduction. In *The Stanford Encyclopedia of Philosophy (Spring 2019 Edition)*, ed. Edward N. Zalta. Stanford University: Metaphysics Research Lab.
van Riel, Raphael. 2014. *The Concept of Reduction*. Cham: Springer.
Verhaegh, Sander. 2017. Blurring Boundaries: Carnap, Quine, and the Internal-External Distinction. *Erkenntnis* 82 (4): 873–890. https://doi.org/10.1007/s10670-016-9848-0.
von Kutschera, Franz. 1991. Carnap und der Physikalismus. *Erkenntnis* 35 (1/3): 305–323.
Walter, Sven. 2006. Multiple Realizability and Reduction: A Defense of the Disjunctive Move. *Metaphysica* 7 (1): 43–65.