Abduction in Philosophy of Mind[*]

Christian J. Feldbacher-Escamilla Maria Sekatskaya

Spring 2026

Abstract

[128] This contribution explores the application of abductive reasoning in the philosophy of mind. It emphasizes the importance of explanatory features and abductive virtues, highlighting the need to balance virtues such as accuracy and simplicity in developing robust explanations. For this purpose, it identifies a list of common desiderata underlying many accounts in the philosophy of mind. While existing discussions of desiderata in the philosophy of mind, such as multiple realizability, mental causation, and qualia, often prioritize individual explanatory features and abductive virtues, this work argues for a more holistic approach that also considers the interplay among these features and virtues.

Kewords. *philosophy of mind*, *abductive virtues*, *multiple realizability*, *mental causation*

7.1 Introduction

Abductive philosophy of mind applies abductive thinking to the philosophy of mind. In this contribution, we argue that this can be done in at least two ways. First, one can use abduction for theory choice by applying abductive criteria to select among alternative theories in the philosophy of mind. Second, one can incorporate abductive features within a theory to address specific problems in the philosophy of mind.

Also, abduction is considered to come in two main forms: selective and creative abduction (cf. Schurz 2008; Feldbacher-Escamilla 2022). Both aim at explanations; creative abduction aims at the construction of explanations, which is more relevant to the second form of abductive philosophy of mind—namely, the use of abductive features within a theory. Selective abduction, on the other hand, aims to select among a set of alternative explanations. [129] This

[*][This text is published under the following bibliographical data: Feldbacher-Escamilla, Christian J. and Sekatskaya, Maria (2026). "Abduction in Philosophy of Mind". In: *Inductive Metaphysics: Insights, Challenges, and Prospects*. Ed. by Hüttemann, Andreas and Schurz, Gerhard. New York: Routledge/Taylor & Francis Group, pp. 128–148. DOI: 10.4324/9781003514404-9. All page numbers of the published text are in square brackets. The final publication is available at https://doi.org/10.4324/9781003514404-9. For more information about the underlying project, please have a look at http://cjf.escamilla.academia.name.]

amounts to an inference to the best explanation and fits the first form of abductive philosophy of mind. Explanatory features and abductive virtues play a key role. However, balancing these features is also crucial, because sometimes we trade one abductive virtue, such as accuracy, for another abductive virtue, such as simplicity, in explanations and predictions. While discussions of explanatory features are already common within the philosophy of mind, their argumentative and epistemic roles are often taken for granted without considering their context. Moreover, existing accounts in the philosophy of mind primarily focus on individual explanatory features rather than on the abductive balancing among them. This contribution outlines how an abductive philosophy of mind can fill this lacuna. Our aim is twofold: first, to study the theoretical background of the discussion of individual abductive virtues as employed within the philosophy of mind. Second, to stress the role of balancing in forming a general perspective on abductive virtues within this field.

We proceed as follows: Section 7.2 briefly outlines the most relevant notions of abduction used in this contribution. Section 7.3 provides a brief overview of discussions within the philosophy of mind, focusing on desiderata serving as cornerstones. Section 7.4 identifies relevant abductive virtues underlying these desiderata. Section 7.5 offers a broader context for considering individual abductive virtues and outlines the role of balancing for a holistic consideration. Section 7.6 concludes our investigation.

7.2 Abduction

Abduction is a form of ampliative reasoning (Schurz 2021, sect.1). It plays a constitutive role in natural science and increasingly in social science and the humanities, including philosophy (cf. Williamson 2016). The theory of abduction traces back to Charles S. Peirce, who was the first to provide a general schema of abductive inferences (cf. Peirce 1994, CP 5.189). Abductive reasoning comes in two forms (cf. Douven 2022; Schurz 2008): First, there is explanatory reasoning in hypothesis generation, or *creative abduction*, which involves formulating theoretical hypotheses containing new concepts or models based on empirical data. As we indicated in the introduction, creative abduction plays an important role in the use of explanatory features within a theory. Second, there is explanatory reasoning in hypothesis justification, or *selective abduction*, which involves selecting among a set of hypotheses and potential explanations. As we indicated, this form of abduction is particularly relevant for the abductive philosophy of mind in selecting among alternative theories in the philosophy of mind. Creative abduction is ampliative because the truth of the premises can never guarantee the truth of a conclusion with *new* relevant concepts. However, it differs from other ampliative inferences, such as induction in the narrow sense, insofar as the latter is about generalizations from the concepts and notions used in the premises, [130] whereas the former introduces new concepts and notions not found in the premises (cf. Schurz 2008, p.202). The prototypical example of creative abduction involves common cause reasoning (cf. Schurz 2008; Feldbacher-Escamilla and Gebharter 2019). Here, an

observed empirical correlation between two events leads to the inference that either one event caused the other or that both events are the result of another, *new* event—a so-called common cause. There are accounts that outline the theoretical relevance of common cause abduction in terms of *unificatory power* (Schurz 2008, for details, cf. and Feldbacher-Escamilla and Gebharter 2019). However, other criteria and virtues, such as *explanatory* and *predictive power* (Schupbach and Sprenger 2011), *simplicity* (Sober 2015), and *loveliness* (Lipton 1991), must also be employed to specify proper creative abductions, in contrast to mere speculative forms of reasoning, such as so-called speculative abductions (cf. Schurz 2008, sect.7.1).

The other form of abduction, selective abduction or inference to the best explanation (Lipton 1991), allows one to select among a set of available alternatives to explain some fact. This comparison involves a trade-off between several epistemic and non-epistemic explanatory virtues. Explanatory virtues are, among others, the *accuracy of an explanation* in terms of its fit with the data, *simplicity*, and *explanatory strength*. These virtues can come into conflict with each other: an explanation might fulfill the simplicity criterion to a very high degree but be less successful in fitting the data, whereas a competing explanation might be less simple but fit the data better. The fulfillment of the criteria must therefore be weighed against each other (cf. Schurz 2008). New accounts within the philosophy of science link abductive reasoning to discussions within model selection and thereby provide epistemological underpinning for abductive virtues such as *unification, explanatory* and *predictive power*, and *simplicity* (ontological and theoretical), due to their role in achieving such epistemic goals as *accuracy* and *expected predictive accuracy* (cf. Forster and Sober 1994; Sober 2015).

This highlights two important roles of abductive reasoning. First, abductive virtues are supposed to have some epistemological underpinning. Thus, arguing on the basis of abductive virtues (within a theory) provides not only conventional or aesthetic reasons but also epistemic ones. Second, abductive virtues (for comparing theories) can be traded off and balanced against each other due to their instrumental character. As is common in accounts of practical rationality—and applicable to rationality in general—whatever kind of trading off or balancing (of means) is instrumental for our epistemic ends is permissible.

We will use both of these roles in our account of an abductive philosophy of mind in Section 7.5. However, before that, we will prepare the ground by sketching relevant discussions within the philosophy of mind. [131]

7.3 Desiderata for Discussions within the Philosophy of Mind

The philosophy of mind encompasses diverse positions, from reductive physicalism, which seeks to explain all mental phenomena in the same way as physical phenomena, through functionalism and non-reductive physicalism, which hold that explanations of mental phenomena differ from those of physical phenomena, even though, metaphysically, the mental depends on the physical, to

different versions of dualism and idealism, which view the mental as both inexplicable by and ontologically distinct from the physical. In this chapter, we focus on versions of physicalism in Kim's (2005) sense of minimal physicalism, analyzing theories of mind that assume that all mental phenomena (M) supervene on physical phenomena (P), where supervenience is defined as a relation between M and P such that there is no M-difference without a corresponding P-difference. The question we want to explore is: what epistemic virtues do these theories rely on to arrive at different answers? We propose to classify these theories based on the following list of competing desiderata (cf. Feldbacher-Escamilla and Sekatskaya 2025, sect.2) that are seen as desirable, though they might not be jointly satisfiable. They are desiderata in the sense that an adequate theory within the philosophy of mind should either account for them, align with them, or offer an error theory explaining why its incompatibility with them is unproblematic or even beneficial.

1. Theoretical unification (cf. Feigl 1958/1967): If we assume supervenience, we assert that there is a systematic correlation between the mental and the physical, even if we don't (yet) know the psycho-physical laws. If the mental can be reductively explained by the physical, theoretical unification is achievable: instead of two types of theories, a physical and a psychological one (or many, if psychology, sociology, economics, etc. are not reducible to each other), we will have one powerful unified theory.

2. Ontological simplicity (cf. Smart 1959): If there is only one type of substance (studied by physics) and only one type of fundamental properties (those described in fundamental physical theories), then our ontology is flat. We may have (infinitely) many individuals, but they are ultimately similar to each other, making the ontology – which describes what there is—ultimately very simple.

3. Autonomy of special sciences (cf. J. A. Fodor 1974): Currently, the special sciences are not reducible to the more fundamental sciences. Even though minimal physicalism assumes that all mental phenomena supervene on physical processes in the brain, there is no unified theory that explains psychology in terms of physics. Indeed, even explaining psychology in terms of neurobiology is far from being achieved. [132]

4. Mental causation (cf. Kim 1998): Our everyday understanding of the world and the special sciences that deal with the mental, such as psychology, sociology, and economics, presuppose that mental states are causally efficacious. This is why explanations in terms of beliefs and desires work so well. If our beliefs and desires weren't causally efficacious, the correlation between our wanting to do something and our bodies' moving accordingly would be utterly mysterious.

5. Multiple realizability (MR; cf. Putnam 1967): Mental phenomena seem to be realizable by different physical systems. We know from experience

that physical processes in our brains give rise to rich subjective experiences: pain and pleasure, tastes, colors, and sounds. We suppose that other animals on Earth also have subjective experiences. It is intuitively plausible that other physical systems, such as extraterrestrial beings and advanced robots with artificial general intelligence, can also have mental states and subjective experiences.

6. Natural kinds (cf. J. A. Fodor 1974): Mental concepts seem to track properties that truly exist in the world. Moreover, these properties are natural in a sense that gerrymandered, accidental, or purely conventional properties are not. Are the properties to which the mental concepts refer purely physical? This seems unlikely, because multiple physical realization bases, presupposed by the MR desideratum, are too diverse and heterogeneous to form one physical kind.

7. Qualia (cf. Chalmers 1996): Qualia are at the center of many discussions about consciousness. They are, according to Chalmers, the "hard problem" of consciousness. It seems that even if we know everything about the physical bases of our conscious states and have the most plausible and detailed specifications of their causal roles, their qualia—or what it is like to be in those states—remain unexplained.

These are central desiderata around which discussions in the philosophy of mind often align. Clearly, many more can be found in the debates, and the ones listed above can easily branch into a broad range of additional desiderata. However, for our general aim of indicating the relevance of particular features of abductive reasoning for the philosophy of mind, this set suffices. In the next section, we will show how different accounts in the philosophy of mind try to meet these different desiderata. In the subsequent section, we will outline a general abductive analysis.

7.4 Explanatory Features within the Philosophy of Mind

In the previous section, we briefly described seven general desiderata that serve as cornerstones of discussions in the philosophy of mind. In this section, we focus on the explanatory features that can be found in these discussions. [133]

Let us begin with the desideratum of ***theoretical unification (1)***. The general discussion of unification is often directly linked to the topic of explanation (cf., e.g. Kneale 1949, p.91; Hempel 1965, p.444; Friedman 1974; Kitcher 1981). The general idea is that unification provides an explanation (because, according to some authors, explanation amounts to unification). On these accounts, unification can be directly considered an explanatory or abductive feature. For an epistemological reconstruction of unification as an explanatory or abductive virtue, see Forster and Sober (1994). However, as Gebharter and Feldbacher-Escamilla (2023) show, common measures of unification fall short of systematically aligning with common measures of explanatory power. Thus, the discus-

sion on successfully linking unification to explanation remains largely qualitative and has not yet achieved quantitative confirmation. In the philosophy of mind, the discussion of the desideratum of theoretical unification does not deviate from these general considerations. We can, therefore, state that in the philosophy of mind, the role of unification is generally considered to be explanatory and, hence, also abductive.

With respect to the second desideratum, ***ontological simplicity (2)***, similar considerations apply to theories about the mind as to other scientific theories (cf. van Riel 2014). The role of ontological simplicity is well-known in the debate about the relevance and justification of "Ockham's razor" (cf. Sober 2015). Newton's methodology famously incorporated it, particularly in relation to causal explanations (cf. Feldbacher-Escamilla 2019). Additionally, an abductive reconstruction of its epistemic merit can be found, e.g., in (Forster and Sober 1994, sect.4). However, the main problem is that, as long as the scientific community has not accepted a full reduction of all special sciences studying the mental to a more fundamental science, strong intuitions remain that no account of the mental has (yet) achieved this desideratum. Thus, the discussion of ontological simplicity is generally linked to explanation and abduction, but it remains at a very general level. As we will see below, more specific discussions can be found in relation to the other desiderata.

Ad ***autonomy (3)***. Special sciences have theories that employ (currently) irreducible mental terms and make empirically testable and successful predictions. Since their predictions work, the mental terms employed in these predictions probably correspond to something real, according to the thesis of scientific realism (List 2019). As long as these mental terms are not reduced to anything more fundamental, we have a prima facie reason to think that what they refer to is not identical to what the terms of more fundamental sciences refer to. As we will discuss in the next section, the relevant abductive feature in this kind of argumentation is *predictive power*.

Ad ***mental causation (4)***. Non-reductive physicalism faces the causal exclusion problem: if mental properties supervene on physical properties but are not identical to them, they cannot be causally efficacious. [134] This is because, according to the thesis of the causal closure of the physical, usually assumed by both reductive and non-reductive physicalists, for every physical event, there is a sufficient physical cause, leaving no causal role for mental properties (Kim 1998, 2005). In other words, mental properties do not influence what happens at the physical level, including our actions. They are either epiphenomenal or, at best, they systematically overdetermine effects that are already caused by physical properties.

One prominent non-reductive physicalist strategy proposed to address the causal exclusion problem is based on considerations of proportionality. According to Yablo (1992), Woodward (2008, 2015), and List and Menzies (2009), mental causes are efficacious and not overdetermined by physical causes because they are more proportionate to their effects than the physical processes on which these mental causes supervene. While physical processes are sufficient to cause their outcomes, they are overly specific. For example, a particular

configuration of a neuronal process in a person's brain is causally sufficient for that person ordering a cappuccino, but the explanation in terms of an intention to order one would be more fitting, because, presumably, this person could have ordered the cappuccino even if the pattern of neuronal excitation in her brain was slightly different. This shows that the mental can be regarded as a cause of behavior. However, whether the mental, rather than the physical, is the cause, or whether the mental can be a cause alongside the physical, remains an open question, explored in the context of "strong" and "weak" proportionality (cf. McDonnell 2017).

Baumgartner (2009) objected that according to an interventionist account of causation and assuming non-reductive supervenience, the mental cannot cause anything because it cannot be manipulated independently of its physical base. Woodward (2008) replied that supervenient M-properties are not in a causal relation to their physical bases P and, hence, not supposed to vary independently of P. The intervention on M happens simultaneously with the intervention on P. Although M cannot causally influence other mental or physical states by itself, it sometimes provides a better, more proportional, explanation. Gebharter (2017) demonstrated the validity of the exclusion argument using causal Bayes nets, assuming that supervenience relations are treated formally in the same way as causal relations. Whether the causal exclusion problem can be solved depends on the account of causation we accept and on whether we treat M-properties differently from P-properties. This, in turn, largely depends on the strength of our reasons to believe in a non-reductive supervenience relation that is not explained by identity. Additionally, it also hinges on our understanding of the nature of the supervenience relation itself. One way to defend the causal efficacy of the mental is to weaken strict supervenience to probabilistic supervenience, which is the thesis that every change in the values of causal variables representing M-properties makes a probabilistic difference for at least one of the values of the variables representing their P-bases (Gebharter and Sekatskaya 2024). [135]

In general, we observe that different notions of causation and proportionality play a central role in the discussion of mental causation. As we will outline in the next section, these notions typically come with abductive reading or abductive implications (concerning causal explanation/prediction and the proportionality of an explanation).

Ad ***multiple realizability (5)***. One of the main reasons to prefer non-reductive physicalism to reductive physicalism, despite the challenge of causal exclusion, is the assumption of MR. The main reactions to the MR desideratum include accepting functionalism, accepting kind-specific identifications of mental states, going disjunctivist, or rejecting MR.

Functionalists argue that the functionalist explanation is both simpler and more general than the identity explanation: it offers generalizations at the *proper level of description* and secures the autonomy of psychology (Putnam 1967; J. A. Fodor 1974). However, if functional properties are not token-identical to the underlying physical properties, as role functionalism claims, it faces the problem of causal exclusion (cf. Kim 1998, 2005). In con-

trast, realizer functionalism avoids this problem by asserting token identity, because mental properties have the same causal powers as their realizing physical bases. Most versions of realizer functionalism define mental properties by their causal roles, which can be realized by different physical systems (Levin 2022; Lewis 1972, 1994). This leads to a medium-level type of identity: either kind-specific or identity with subsets of causal powers. If mental properties are identified with a specific kind of physical system realizing a given function – such as the pain function being realized by different kind-specific physical systems in humans, octopi, and extra-terrestrials—then there is no single functional kind "pain". Instead, there are more fine-grained functional kinds of pain in different physical systems (Kim 1998, 2005). The downside of this approach is that it loses some of the generality of role functionalism and risks becoming indistinguishable from identity theory if, even within a single kind, individuals differ so significantly that no single type of physical system performing a given function is common to all. Alternatively, functions can be identified with subsets of causal powers of the physical systems realizing them (Wilson 2011). For example, only those properties of the human and non-human neural systems that are necessary for performing a pain function will be included in that subset. This secures the explanatory role of functions. Whether these subsets of functions are similar across all systems remains an open question. Disjunctivists propose a different metaphysical route to save the reality of mental kinds. They claim that each predicate referring to a mental kind is co-referential with a disjunction of physical predicates ($P_1 \vee P_2 \vee \cdots \vee P_n$), so that the mental kind "pain" is identical to the physical kind "pain in humans *or* pain in octopi *or* pain in extra-terrestrials *or* …" (Clapp 2001; Walter 2006). The arguments against this move are based on *explanatory* considerations of naturalness and projectability and will be discussed shortly. [136]

Finally, the thesis of MR itself can be rejected. Polger and Shapiro (2016) argue that to pose a challenge to identity theory, MR must meet very specific criteria. It must be more than a mere variation within a single kind; otherwise, it becomes trivialized, since all scientifically respectable and successfully reduced kinds, such as temperature, can be realized in more than one way (e.g., temperature in gases vs temperature in solids). At the same time, a realized function must be very similar, or even identical, across realizations, because if there are functionally relevant physical differences between different realizations, a better *explanation* might be to postulate two similar kinds, rather than one multiply realized kind. According to Polger and Shapiro (2016), we do not have empirical support for this type of MR. All mental functions that we are currently familiar with occur in living creatures on Earth, whose neural systems are composed of the same micro-constituents organized in similar ways. MR is intuitively plausible as a modal thesis: it seems to us that mental properties can be realized by very different physical systems. But how much should we trust this intuition when deriving metaphysical conclusions? Again, explanatory and abductive features play a central role in the discussion of this desideratum. One is the *proper level of description*, which resembles a form of *proportionality* discussed in relation to mental causation (4). Another concerns

naturalness and *projectability*, as will be discussed below, when we address natural kinds (6).

Ad ***natural kinds (6)***. Some mental properties seem to constitute natural kinds, enabling generalizations using mental predicates in the special sciences. According to (J. Fodor 1997; J. A. Fodor 1974), disjunctive designators cannot pick out natural kinds, drawing on reasoning made famous by Armstrong's (1978) "x is a raven $\vee$ x is a writing desk" example. However, there is a problem with postulating that mental kinds (M) are natural while their physical instantiation bases (P) do not form a natural kind. If a predicate that refers to a mental kind M is co-referential with a disjunction of physical predicates ($P_1 \vee P_2 \vee \cdots \vee P_n$), then either both of these predicates refer to one and the same kind in the actual world, or both fail to refer to it (Kim 1992, p.15). One can argue that mental and physical predicates still pick out different kinds in the actual world if, in some other possible worlds, mental predicates refer to non-physical properties. Whether physicalists should take this possibility seriously is a separate question (cf. Stoljar 2010).

Another reason to disqualify disjunctive predicates is that psycho-physical laws containing disjunctive predicates are not confirmed by their positive instances and thus *unprojectable*, in the same way as the generalization "All jade is green" is not a law, since jade is either jadeite or nephrite, and these two minerals have different micro-structure and different causal powers (Kim 1992, pp.11–13, 1998, pp.106–110; Walter 2006, ppp.56–58). However, disjunctive designators can be causally homogenous "if all *and only* the individuals satisfying them have something in common" (Walter 2006, p.59). Therefore, if mental properties are defined by their causal roles, [137] in line with what many functionalists claim, and in particular if a causal profile of a certain mental property M is the subset of the causal powers of the set of all the causal powers of its physical realizers P in virtue of which these P have this causally defined mental property M, then all disjuncts of a disjunctive physical designator of M have something in common, namely, they refer to this subset of causal powers (Walter 2006, pp.59–63). This line of defense, of course, presupposes that mental properties are individuated by their causal roles and that there is something causally similar between all physical realizations of these mental properties that have the same causal powers (as we discussed above, this is how the MR thesis can be interpreted). If mental properties cannot be causally defined, as argued by some proponents of non-reductive physicalism, particularly defenders of the irreducibility of qualia, it becomes unclear whether all physical realizers of a given qualia have something causally homogeneous in common. The main abductive virtue in this debate is *projectability*: natural kinds allow for a projection of their instances, whereas purely conventional kinds do not.

Ad ***Qualia (7)***. Qualia are usually not defined but introduced by example. They are the way that something looks, feels, or tastes to us, like a specific shade of red of a ripe tomato, the warmth of the summer sun on our skin, or the aroma of freshly ground coffee. Non-reductive physicalists have a strong intuition that qualia cannot be reductively explained: even if we know everything about their causal roles and physical bases, and even if we have a com-

plete mapping of all qualia states to the corresponding physical and functional states, along with *explanations* of how the causal mechanisms of these physical states produce these functional states, there is still the feeling that something is left *unexplained*—namely, why do these states feel like *this*? This is the intuition illustrated by the knowledge argument (Jackson 1982), the explanatory gap argument (Levine 1983), and the possibility of phenomenal zombies (Chalmers 1996). Functionalists and reductive physicalists have offered various analyses of qualia, supposedly showing that they can be functionally *explained* (cf. Dennett 1991; Lewis 1994; Levin 2002; Shoemaker 2007), or at least fit into a physicalist worldview by means of identity with certain brain states without any problems (cf. Loar 1990; Place 2004; Polger 2011). In this contribution, we cannot delve into a detailed analysis of different accounts of qualia, but we can briefly overview what follows from the widely accepted status of qualia as a desideratum.

First, *reality and special status*: there is a strong intuition that qualia are real and known to us in a very special, intimate, and subjective way, distinct from the usual way we know other phenomena. Any theory that aims to *explain* why this is not the case faces an uphill battle. This is why the elimination of qualia is not a popular strategy, and the explanation of them in purely functional terms is seen as problematic. While this is not necessarily a decisive reason to reject functionalism, it is at least a serious downside of this theory. [138] Friends of qualia think that qualia are a *datum to be explained, not explained away*. They either opt for some form of identity theory (type or token, depending on their views about reduction—for an overview of different positions on reduction in debates in the philosophy of mind, see Feldbacher-Escamilla and Sekatskaya 2025), or they make an exception for qualia in an otherwise well-explained physicalist worldview (Kim 2005; Stoljar 2010).

Second, *causal (in)efficacy*: Functionalists argue that if qualia are real, they must have some causal influence. To the extent that they have this influence, they can be functionally analyzed. In response, proponents of qualia can either deny their causal efficacy and accept epiphenomenalism (Jackson 1982; Robinson 2018), or argue that, even though qualia do have causal influence, there is more to them than their causal powers—a "more" that resists any analysis in non-experiential terms. In our brief overview, we have already marked the relevant parts of abductive reasoning for this desideratum: it is directly about the *explanation* of qualia.

In the following section, we want to systematically combine our discussion of abductive features and outline their interrelation in the debate about the individual desiderata.

7.5 Abductive Philosophy of Mind

As discussed in Section 7.2, abductive reasoning has two important components. The first is based on the (epistemic) merit of individual abductive virtues, such as accuracy, unification, and simplicity, in arguing for an account or explanation. The second adopts a more holistic perspective by balancing

and weighing between abductive virtues for or against an account or explanation.

Let us begin with the first component. In Section 7.3, we outlined the key cornerstones of discussions within the philosophy of mind, resulting in a list of seven desiderata. In Section 7.4, we focused on the abductive virtues prominent in the discussion of these desiderata. Our conclusion was that the desiderata of ***unification (1)*** and ***(ontological) simplicity (2)*** are themselves abductive virtues. The main abductive virtue in discussions about the ***autonomy of the special sciences (3)*** is *predictive power*. That in discussions about ***mental causation (4)*** is *proportionality*. In discussions of ***multiple realizability (5)***, two abductive virtues are important. The first is the same as in (4), i.e., proportionality in terms of demanding to use the "right level of description". The second is the same as in the discussion of ***natural kinds (6)***, namely, *projectability*. Finally, the main abductive virtue involved in discussions of ***qualia (7)*** is *explanatory power*.

Let us begin with predictive power as it is used in discussions of ***autonomy (3)***. In this context, predictive power—an abductive virtue [139]—is considered a feature that supports non-reductive accounts within the philosophy of mind. Now, in which sense is predictive power an abductive virtue? Broadly speaking, there are two approaches to incorporating predictive power into the framework of abduction. First, there is a positive account of the thesis of structural identity of explanation and prediction (for a discussion, see Hempel 1965, pp.366-376). The idea is that if prediction and explanation are structurally identical, i.e., if explaining some *explanandum* based on some *explanans* is logically equivalent to predicting (or retrodicting) the explanandum based on the explanans, then it is plausible to assume that if features of explanations constitute abductive virtues for trivial reasons (by definition), then features of predictions also constitute abductive virtues. However, the structural identity thesis is contested. For this reason, it is important to stress the second line of reasoning in favor of predictive power as an abductive virtue. As we indicated in Section 2, the epistemic justification of a central abductive virtue (namely *simplicity*) is sometimes discussed in the context of model selection, where it is shown that simplicity epistemically matters for the *expected* performance of a model. It matters inasmuch as simplicity can be instrumental for the expected *predictive accuracy* of an account, which is an essential component of an account's *predictive power*. With regard to the application of abductive methodology in statistics, the shift from *explanatory* power in terms of the accuracy of an explanation towards *expected* explanatory power, i.e., *predictive* power, can be clarified by differentiating between descriptive statistics, which covers explanation, vs inferential statistics, which covers prediction (cf. Otsuka 2023). Hence, in abductive reasoning (and statistics), predictive power is not only *an* abductive virtue but even a core virtue, allowing for an instrumental justification of other central abductive virtues, such as simplicity. If non-reductive physicalism shares this abductive virtue and reductive physicalism falls short of it, then the individual consideration of the abductive virtue *predictive power* clearly favors non-reductive physicalist accounts in the philosophy of mind.

Regarding ***mental causation (4)*** and its main abductive virtue, *proportionality*, the analysis is straightforward, since relevant authors in this field directly emphasize the role of proportionality in explanation. For example, List and Menzies (2009) have an account of proportional difference-makers and stress that "the notion of causal relevance or difference-making plays a central role in theories of causal *explanation*" (sect.4f). Yablo (1992) makes a similar argument. Thus, proportionality is considered an explanatorily relevant virtue, hence also an abductive virtue. Taken individually, it is seen as favoring non-reductive physicalist accounts of mental causation over their reductive physicalist rivals. [140] We should mention that our brief discussion of proportionality focuses on accounts that advocate for "strong proportionality", which posits that there is only one cause in play (cf. McDonnell 2017, sect.5.1). However, for abductive considerations, our discussion may also be generalized, and it could suffice to rely only on "weak proportionality", which states that there might be several causes in play, but some of them are more optimal than others (we thank an anonymous reviewer for bringing this distinction to our attention).

Regarding ***multiple realizability (5)***, we have already noted that the discussion of abductive virtues is similar to that of mental causation (concerning the *proper level of description*) and what will follow below regarding *naturalness and projectability*. A specificity of the debate on MR should also be mentioned here. As discussed in Section 7.4, two significant accounts aim at fulfilling the desideratum of MR. On the one hand, functionalism argues that a functional description alone suffices for prediction, and thus there is no need to know the exact physical realizers of a mental property M to make an adequate prediction about M. Therefore, functions serve as *unifiers* here. On the other hand, disjunctivism accounts for MR by disjunctively considering possible physical realizers. Thus, functionalism and disjunctivism take completely opposing approaches to the role of realizers. These opposing accounts involve trade-offs concerning abductive virtues, as we will discuss below.

We have identified the abductive virtue of *projectability* as relevant for the discussion of the desideratum of ***natural kinds (6)***. A property is projectable if it is instrumental for the success of inductive reasoning and non-projectable otherwise. Success of induction is a particular form of *predictive power*, namely predictive power due to inductive reasoning. In this sense, projectability is instrumental for predictive power and, thus, qualifies as an abductive virtue. The context of natural kinds is a classical one (cf. Bird 2018) but not the only one. Goodman (1955/1983), e.g., argues for projectability in terms of the rightness of categorizations via entrenchment relations; for discovering natural kinds, not directly projectability but entrenchment is relevant.

Turning to the final desideratum, ***qualia (7)***, the relevant abductive virtue, *explanatory power*, is straightforward. In Section 7.4, we observed that functionalism struggles to account for this desideratum. However, while most functionalist accounts encounter this issue, certain mixed forms of functionalism appear capable of circumventing it by employing abductive reasoning. Here, we outline a functionalist reconstruction of kind-specific reductions, based on Kim (cf. 1992). We begin by observing that different people and non-human an-

imals exhibit similar behaviors. Since such behaviors are accompanied by subjective feelings/qualia in our own experience, by induction we also attribute similar qualia to other creatures. [141] Postulating similar qualia for different physical systems helps to predict similar behaviors, thus justifying this inductive generalization through abduction. According to functionalism, pain plays a sufficiently similar functional role in humans and in other creatures (such as non-human animals, Martians, etc.). When we ask ourselves what explains the similar behavior among them, the answer can lie in their similar qualia—without requiring a complete functional analysis of, for instance, what pain is for every possible creature. Such an analysis would be highly demanding and might take a form logically similar to disjunctivism: "if you are human, then the functional role is such and such; if you are a Martian, then the functional role is such and such; if ...". Rather than attempting to define an overly complex general function, abductively postulating similar qualia across species provides explanations and predictions that are theoretically simpler. Thus, while functionalism generally struggles to meet this desideratum, abductively adapted forms of functionalism—those that infer not only functions but also qualia – might be able to satisfy it, if one is willing to concede that a precise functional analysis of mental properties in terms of the causal roles of physical properties is not always possible.

This concludes the individual discussion of abductive virtues. Regarding the holistic component, in Section 7.4, we outlined that satisfying one desideratum often introduces challenges for others. Though this undermines the hypothesis that all desiderata can be jointly satisfied, it does not diminish their relevance, as abductive virtues play a key role within each desideratum. Moreover, as we emphasized in Section 7.2, trading off and balancing among abductive virtues is a standard procedure in abduction. As shown in the discussion of abductive reasoning in model selection, this balancing is not a shortcoming of abductive reasoning, but rather one of its strengths. Due to space constraints, we can only briefly and selectively outline this idea here. We see several relations between the desiderata stated in Section 7.3 as relevant for balancing among abductive virtues. For example, ***multiple realizability (5)*** partly depends on *proportionality* considerations from the context of ***mental causation (4)*** and *projectability* considerations from the context of ***natural kinds (6)***. The desideratum regarding ***qualia (7)*** partly conflicts with that of ***mental causation (4)***, as exemplified in epiphenomenalist accounts that argue for the causal inertness of qualia. In principle, it should be straightforward to find, for each pairwise comparison of desiderata, positions within the philosophy of mind that either positively address them by establishing a dependence between the desiderata or at least a joint (if not mutual) satisfaction, or negatively address them by systematically failing to account for one desideratum while succeeding with the other. A further, more problematic possibility—not discussed here—is a systematic failure to satisfy both desiderata. For illustrative purposes, we will discuss two cases of abductive balancing. [142]

Let us begin with the positive case of satisfying two desiderata due to their mutual dependence. From an abductive perspective, such dependence can

be identified, e.g., between ***autonomy (3)*** and ***mental causation (4)***: The discussion of ***(4)*** is directly linked to that of ***(3)***. If ***(4)*** is based on a narrow notion of *causation* and a "strong" version of *physicalism* (i.e., deterministic rather than probabilistic supervenience), then with the causal inertness of the mental comes also the subordination of the special sciences to foundational sciences such as physics. However, if ***(4)*** is based on a wider notion of *causation* as discussed, e.g., in different proportionality accounts of causation and explanation, or if ***(4)*** is based on a weaker version of *physicalism* (probabilistic), then the autonomy of the special sciences is justified. Our discussion was only about one direction of the link, from ***(4)*** to ***(3)***. However, one can also argue for the link in the opposite direction, from ***(3)*** to ***(4)***. From the success of the special sciences in terms of predictive (but also explanatory) power concerning the mental (M), as opposed to the current failure of foundational sciences such as physics (based on P) to predict and explain M, one can infer the reality of M, providing not only an argument for the ***autonomy*** of the special sciences ***(3)*** but also for the role of notions such as *proportionality* as we find in ***(4)***; by this kind of reasoning, such notions gain a realistic underpinning. Given this argument, proportionality is not only an epistemic feature of explanations but also has a metaphysical/ontological foundation. It is a *real* feature, not just one *constructed* for our explanations. One might wonder how this inference from the epistemic success of the special sciences towards a *real* or ontological basis of their kinds and laws can be justified. Once again, abductive reasoning is well-suited to justify this inference: assuming the existence of the kinds and regularities of the special sciences is what *best explains* their success (for a more general discussion of abduction as a method for metaphysical inferences cf. Schurz 2021). This also concludes our outline of the bidirectional dependence between (3) and (4) based on abductive reasoning.

An example of negative dependence can be identified with respect to ***multiple realizability (5)***, ***natural kinds (6)***, and ***qualia (7)***. Disjunctivism and some forms of functionalism face this problem: accounting for a diverse set of possible realizers often results in unnatural or non-projectable properties (as is the case with disjunctivism). Conversely, focusing on natural and projectable properties tends to pose challenges in explaining a diverse set of possible realizers (as is the case with some forms of functionalism). The same applies to ***(5)*** and ***(7)***, although mixed forms of functionalism that abductively infer not only the functional roles of mental properties but also their qualia might be able to circumvent this negative dependence. In our discussion of ***multiple realizability (5)***, we noted the opposing relationship between functionalism and disjunctivism regarding the physical realizers (P) of mental properties. [143] In particular, functionalism faces problems of *explanatory power* with respect to explaining qualia (see our brief discussion of ***(7)***). On the other hand, it gains in terms of *unification* in the context of ***multiple realizability (5)***. Hence, it seems to trade its *unificatory* gains with respect to ***multiple realizability (5)*** for its failure regarding *explanatory power* concerning ***qualia (7)***. Disjunctivism, on the other hand, trades its gains in *explanatory power* in the context of ***multiple realizability (5)*** for its failure to account for naturalness and, in partic-

ular, *projectability* (6). One can also note that disjunctivism gains with respect to ***ontological simplicity*** (2) by focusing only on realizers and not introducing functions. It does so, however, at the cost of failing to *unify*, because its explanations are theoretically complex, arguing case by case (P-realizer$_1 \vee$ P-realizer$_2 \vee$ P-realizer$_3$), etc. This indicates that trading off and balancing among abductive virtues is relevant for discussions and accounts within the philosophy of mind.

Whether our brief discussion in this contribution provides a unifying thread in terms of abduction is an important question of abductive philosophy of mind. The same applies to whether some of the abductive virtues discussed here favor, e.g., non-reductive over reductive physicalist accounts. Since this chapter outlines a program and not a particular account of abductive philosophy of mind, these questions cannot be addressed here, but they remain open for future research.

7.6 Conclusion

We hope this brief discussion illustrates our idea that trading off and balancing are common and necessary for a joint consideration of desiderata within the philosophy of mind. Focusing on abductive virtues relevant to the individual desiderata naturally brings in the possibility of systematically trading off and balancing among them. For this reason, we believe that philosophy of mind can benefit from incorporating abductive considerations more explicitly into its discussions, thereby contributing to the development of an *abductive philosophy of mind*.

Funding and Acknowledgments

This work was supported by the DFG (Deutsche Forschungsgemeinschaft), research unit FOR 2495, research grant SCHU 1566/11–2. In addition, Christian J. Feldbacher-Escamilla would like to acknowledge the support of the Japan Society for the Promotion of Science (JSPS, 2024, PE24725). For valuable comments, we are indebted to Alexander Gebharter, Vera Hoffmann-Kolss, Andreas Hüttemann, Jan Michel, Raphael van Riel, Gerhard Schurz, and Corina Strössner. [144]

References

Armstrong, David M. (1978). *Universals and Scientific Realism*. Cambridge: Cambridge University Press.

Baumgartner, Michael (2009). "Interventionist Causal Exclusion and Non-reductive Physicalism". In: *International Studies in the Philosophy of Science* 23.2, pp. 161–178. DOI: 10.1080/02698590903006909.

Bird, Alexander (2018). "The Metaphysics of Natural Kinds". In: *Synthese* 195.4, pp. 1397–1426. DOI: 10.1007/s11229-015-0833-y.

Chalmers, David (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press.
Clapp, Lenny (2001). "Disjunctive Properties. Multiple Realizations". In: *The Journal of Philosophy* 98.3, pp. 111–136.
Dennett, Daniel C. (1991). *Consciousness Explained*. Boston: Back Bay Books.
Douven, Igor (2022). *The Art of Abduction*. Cambridge, Massachusetts: MIT Press. DOI: 10.7551/mitpress/14179.001.0001.
Feigl, Herbert (1958/1967). *The 'Mental' and the 'Physical': The Essay and the Postscript*. Minneapolis: University of Minnesota Press.
Feldbacher-Escamilla, Christian J. (2019). "Newtons Methodologie: Eine Kritik an Duhem, Feyerabend und Lakatos". In: *Archiv für Geschichte der Philosophie* 101.4, pp. 584–615. DOI: 10.1515/agph-2019-4004.
— (2022). "Meta-Abduction. Inference to the probabilistically best prediction". In: *Philosophy of Computing*. Ed. by Lundgren, Björn and Nuñez Hernandez, Nancy Abigail. Philosophical Studies Series. Cham: Springer Nature, pp. 51–72. DOI: 10.1007/978-3-030-75267-5_2.
Feldbacher-Escamilla, Christian J. and Gebharter, Alexander (2019). "Modeling Creative Abduction Bayesian Style". In: *European Journal for Philosophy of Science* 9.1, pp. 1–15. DOI: 10.1007/s13194-018-0234-4.
Feldbacher-Escamilla, Christian J. and Sekatskaya, Maria (2025). "Reductionism, Supervenience, and Carnap's Account of Empirical Confirmability". In: *Journal for General Philosophy of Science* 56.3, pp. 345–371. DOI: 10.1007/s10838-025-09728-6.
Fodor, Jerry (1997). "Special Sciences: Still Autonomous after All these Years". In: *Noûs* 31.s11, pp. 149–163. DOI: 10.1111/0029-4624.31.s11.7.
Fodor, Jerry A. (1974). "Special Sciences (or: The disunity of science as a working hypothesis)". In: *Synthese* 28.2, pp. 97–115. DOI: 10.1007/BF00485230.
Forster, Malcolm R. and Sober, Elliott (1994). "How to Tell When Simpler, More Unified, or Less Ad Hoc Theories Will Provide More Accurate Predictions". In: *The British Journal for the Philosophy of Science* 45.1, pp. 1–35. DOI: 10.1093/bjps/45.1.1.
Friedman, Michael (1974). "Explanation and Scientific Understanding". In: *The Journal of Philosophy* 71.1, pp. 5–19. DOI: 10.2307/2024924.
Gebharter, Alexander (2017). "Causal Exclusion and Causal Bayes Nets". In: *Philosophy and Phenomenological Research* 95.2, pp. 353–375. DOI: 10.1111/phpr.12247.
Gebharter, Alexander and Feldbacher-Escamilla, Christian J. (2023). "Unification and Explanation from a Causal Perspective". In: *Studies in History and Philosophy of Science* 99, pp. 28–36. DOI: 10.1016/j.shpsa.2022.12.005.
Gebharter, Alexander and Sekatskaya, Maria (2024). "Mental Causation, Interventionism, and Probabilistic Supervenience". In: *Synthese* 203.206, pp. 1–18. DOI: 10.1007/s11229-024-04608-w.
Goodman, Nelson (1955/1983). *Fact, Fiction, and Forecast*. Ed. by Putnam, Hilary. Fourth Edition. Harvard: Harvard University Press.
Hempel, Carl G. (1965). *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. New York: Free Press.

Jackson, Frank (1982). "Epiphenomenal Qualia". In: *The Philosophical Quarterly* 32.127, pp. 127–136. DOI: 10.2307/2960077.
Kim, Jaegwon (1992). "Multiple Realization and the Metaphysics of Reduction". In: *Philosophy and Phenomenological Research* 52.1, pp. 1–26. DOI: 10.2307/2107741.
— (1998). *Mind in a Physical World: An essay on the mind-body problem and mental causation*. Cambridge, MA: MIT Press.
— (2005). *Physicalism, or Something Near Enough*. Princeton: Princeton University Press.
Kitcher, Philip (1981). "Explanatory Unification". In: *Philosophy of Science* 48.4, pp. 507–531. DOI: 10.1086/289019.
Kneale, Matthew (1949). *Probability and Induction*. Oxford: Oxford University Press.
Levin, Janet (2002). "Is Conceptual Analysis Needed for the Reduction of Qualitative States?" In: *Philosophy and Phenomenological Research* 64.3, pp. 571–591. DOI: 10.1111/j.1933-1592.2002.tb00161.x.
— (2022). *The Metaphysics of Mind*. Cambridge: Cambridge University Press.
Levine, Joseph (1983). "Materialism and Qualia: the explanatory gap". In: *Pacific Philosophical Quarterly* 64, pp. 354–361.
Lewis, David K. (1972). "Psychophysical and Theoretical Identifications". In: *Australasian Journal of Philosophy* 50.3, pp. 249–258. DOI: 10.1080/00048407212341301.
— (1994). "Reduction of Mind". In: *A Companion to Philosophy of Mind*. Ed. by Guttenplan, Samuel. Oxford: Blackwell Publishers, pp. 412–431.
Lipton, Peter (1991). *Inference to the Best Explanation*. London: Routledge.
List, Christian (2019). *Why Free Will is Real*. Cambridge, MA: Harvard University Press.
List, Christian and Menzies, Peter (2009). "Nonreductive Physicalism and the Limits of the Exclusion Principle". In: *The Journal of Philosophy* 106.9, pp. 475–502. DOI: 10.5840/jphil2009106936.
Loar, Brian (1990). "Phenomenal States". In: *Philosophical Perspectives* 4, pp. 81–108. DOI: 10.2307/2214188.
McDonnell, Neil (2017). "Causal Exclusion and the Limits of Proportionality". In: *Philosophical Studies* 174.6, pp. 1459–1474. DOI: 10.1007/s11098-016-0767-3.
Otsuka, Jun (2023). *Thinking about Statistics. The Philosophical Foundations*. New York: Routledge. DOI: 10.4324/9781003319061.
Peirce, Charles S. (1994). "Pragmatism and Abduction". In: *Collected Papers of Charles Sanders Peirce*. Ed. by Hartshorne, Charles, Weiss, Paul, and Burks, Arthur W. Cambridge, MA: Harvard University Press.
Place, U.T. (2004). *Identifying the Mind*. Oxford: Oxford University Press.
Polger, Thomas W. (2011). "Are Sensations Still Brain Processes?" In: *Philosophical Psychology* 24.1, pp. 1–21. DOI: 10.1080/09515089.2010.533263.
Polger, Thomas W. and Shapiro, Lawrence A. (2016). *The Multiple Realization Book*. Oxford: Oxford University Press.

Putnam, Hilary (1967). "Psychological Predicates". In: *Art, Mind, and Religion*. Ed. by Capitan, W.H. and Merrill, D.D. Pittsburgh: University of Pittsburgh Press, pp. 37–48.

Robinson, William S. (2018). *Epiphenomenal Mind. An Integrated Outlook on Sensations, Beliefs, and Pleasure*. New York: Routledge. DOI: 10.4324/9780429435348.

Schupbach, Jonah N. and Sprenger, Jan (2011). "The Logic of Explanatory Power". In: *Philosophy of Science* 78.1, pp. 105–127. DOI: 10.1086/658111.

Schurz, Gerhard (2008). "Patterns of Abduction". English. In: *Synthese* 164.2, pp. 201–234. DOI: 10.1007/s11229-007-9223-4.

— (2021). "Abduction as a Method of Inductive Metaphysics". In: *Grazer Philosophische Studien* 98.1, pp. 50–74. DOI: 10.1163/18756735-000098.

Shoemaker, Sydney (2007). *Physical Realization*. Oxford: Oxford University Press.

Smart, J.J.C. (1959). "Sensations and Brain Processes". In: *The Philosophical Review* 68.2, pp. 141–156. DOI: 10.2307/2182164.

Sober, Elliott (2015). *Ockham's Razors. A User's Manual*. Cambridge: Cambridge University Press.

Stoljar, Daniel (2010). *Physicalism*. London: Routledge.

van Riel, Raphael (2014). *The Concept of Reduction*. Cham: Springer.

Walter, Sven (2006). "Multiple Realizability and Reduction: A Defense of the Disjunctive Move". In: *Metaphysica* 7.1, pp. 43–65.

Williamson, Timothy (2016). "Abductive Philosophy". In: *The Philosophical Forum* 47.3-4, pp. 263–280. DOI: 10.1111/phil.12122.

Woodward, James (2008). "Mental Causation and Neural Mechanisms". In: *Being Reduced: New Essays on Reduction, Explanation, and Causation*. Oxford: Oxford University Press, pp. 218–262.

— (2015). "Interventionism and Causal Exclusion". In: *Philosophy and Phenomenological Research* 91.2, pp. 303–347. DOI: 10.1111/phpr.12095.

Yablo, Stephen (1992). "Mental Causation". In: *Philosophical Review* 101.2, pp. 245–280. DOI: 10.2307/2185535.