Synthese
https://doi.org/10.1007/s11229-021-03170-z

ORIGINAL RESEARCH

The distinct existences argument revisited

Wolfgang Barz[1]

Received: 5 June 2020 / Accepted: 20 April 2021

Abstract
The aim of this paper is to take a fresh look at a discussion about the distinct existences argument that took place between David Armstrong and Frank Jackson more than 50 years ago. I will try to show that Armstrong's argument can be successfully defended against Jackson's objections (albeit at the price of certain concessions concerning Armstrong's view on the meaning of psychological terms as well as his conception of universals). Focusing on two counterexamples that Jackson put forward against Hume's principle (which is central to Armstrong's argument), I will argue that they are either compatible with Hume's principle, or imply a false claim. I will also look at several other considerations that go against Hume's principle, such as, for example, Kripke's origin essentialism and counterexamples from aposteriori necessity.

Keywords Armstrong · Jackson · Incorrigibility · Hume's principle · Necessity · Natural laws

1 Introduction

It has now been more than 50 years since the publication of David Armstrong's *A Materialist Theory of the Mind*. Few books have made such a lasting impact on the philosophy of mind, and many of the issues it touches on are still today the subject of lively debate.[1] One of these issues is Armstrong's so-called distinct existences argument. According to this argument, a mental state φ is distinct from the belief that one is in φ; thus, it is possible that φ occurs without the belief that one is in φ, and vice versa.[2]

[1] See Anstey and Braddon-Mitchell (forthcoming) for an appreciation of the relevance of Armstrong's book.

[2] Cf. Armstrong (1968, p. 106). The argument had already appeared in Armstrong (1963, p. 422).

Wolfgang Barz
wbarz@zedat.fu-berlin.de

[1] Department of Philosophy, Goethe-University, Norbert-Wollheim-Platz 1,
60629 Frankfurt am Main, Germany

Published online: 30 April 2021

The context in which Armstrong originally formulated his argument concerned the question of whether it is possible to be mistaken about one's own current conscious experiences. Philosophers sympathetic to dualism typically sought to defend the infallibility thesis, while philosophers who, like Armstrong, saw themselves as materialists, sought to undermine it.[3] However, Armstrong's argument never succeeded in securing the support of the majority of philosophers; and consequently, the idea of infallibility has persisted. As a survey of the literature reveals, Frank Jackson's "Is There a Good Argument Against the Incorrigibility Thesis?" has played no small part in this process.[4] Even those who were skeptical about the idea of infallibility conceded that Jackson's counterarguments were quite ingenious.[5] Thus, it seems that those who would argue against infallibility by appealing to the distinct existences argument would run into difficulties. In my view, this development is a historical injustice. As I will show in the following, Jackson's counterarguments are not as good as they are generally assumed to be.

Revisiting the distinct existences argument is also philosophically relevant for another reason. In the wake of Shoemaker's argument against self-blindness (which initially was intended as an argument against Armstrong's view on introspection),[6] there has been a flowering of so-called constitutive accounts of self-knowledge in recent years, according to which in order to believe that one is in mental state φ, it is already sufficient to be in φ (given certain background conditions such as rationality, possession of relevant concepts, etc.).[7] If the distinct existences argument is conclusive, constitutive accounts lose much of their attraction.[8] Thus, the question whether

[3] Armstrong (1968, p. 103). That the infallibility thesis is incompatible with materialism is not generally accepted. See Pappas (1976) for a critique and Armstrong (1976) for a reply. Lewis (1972, p. 258) attempts to show that no difficulty arises for his version of materialism (which is very similar to Armstrong's) from the assumption that the infallibility thesis is true. The idea, roughly speaking, is to assume that among the folk-psychological platitudes that define psychological terms are some to the effect that a belief that one is in pain never occurs unless pain occurs. To be sure, this does not exclude the possibility that the material realizer of the belief role (a particular neural state N_1) may sometimes occur without the neurological realizer of the pain role (another neural state N_2) occurring. But in this case, N_1 does not count as a belief that one is in pain, and nor does N_2 then count as pain. Thanks to an anonymous reviewer for bringing that passage of Lewis' paper to my attention. For recent examples of philosophers with dualistic sympathies who defend a limited version of the infallibility thesis see Chalmers (2003), Gertler (2012) and Horgan and Kriegel (2007).

[4] Jackson (1973, especially pp. 51–53 and 57–59). A few years previously, Charles Raff had published a critique of Armstrong (1963) that already anticipated some of Jackson's counterarguments; see Raff (1966).

[5] See, for example, Ellis (1976, p. 116): "I cannot in fact find any serious flaw in Jackson's refutations of Armstrong … and others who have tried to argue that I.C.T. i.e. the incorrigibility thesis is false." This tradition of praising Jackson's paper is perceptible even in writings from the recent past. Cf. Gallois (1996, p. 21). A notable exception is Verges (1974), who defends two specific attacks on the incorrigibility thesis—"the classification argument" and "the argument from science"—against Jackson's counterarguments. However, neither the classification argument nor the argument from science is Armstrong's.

[6] Shoemaker (1994)

[7] Cf. Bilgrami (2006), Boyle (2011), Coliva (2009), Heal (2002) and Zimmerman (2006).

[8] I am being deliberately cautious here. For, of course, the fact that the distinct existences argument succeeds does not automatically mean that constitutive accounts of self-knowledge are wrong. Allow me to make two remarks on that. First: proponents of the constitutive account do not simply claim that being in a mental state φ is sufficient for believing that one is in φ. Rather, they typically claim that being in φ, *given that certain background conditions obtain*, is sufficient for believing that one is in φ. However,

it is possible that I am in mental state φ without believing that I am in φ (and vice versa) has fundamental significance for both the ontology and epistemology of mental states.

I will proceed as follows: in Sect. 2, I will first take a closer look at the logical structure of the distinct existences argument. In particular, I will establish what exactly is meant by Hume's principle. It will become apparent that we have to read Hume's principle *de re* if it is to be at all plausible. In Sect. 3, I will discuss Jackson's first counterexample concerning the allegedly necessary connection between husbands and wives. In Sect. 4, I will discuss a counterexample based on Kripke's origin essentialism. In Sects. 5 and 6, I will take a look at Jackson's second counterexample, which refers to the allegedly necessary connection between the properties of being colored and being extended. In Sect. 7, I will examine counterexamples based on aposteriori necessities. None of these counterexamples is successful—or so I will argue.

Before I turn to the details of the distinct existences argument, let me state that my purpose here is neither to defend Armstrong's specific materialist conception of mental states, nor to defend his theory of universals. Rather, I wish to evaluate the distinct existences argument from a neutral standpoint, incorporating as few substantive assumptions about the nature of mental states or universals as possible. Thus, I do not feel committed to Armstrong's claims about the nature of mental states or universals in my argument, and the fact that I may formulate views on those matters that are incompatible with Armstrong's views should not be seen as detrimental to my considerations.

2 Armstrong's distinct existences argument

Armstrong's distinct existences argument appears in both "Is Introspective Knowledge Incorrigible?" and *A Materialist Theory of the Mind*. In my view, the version which appears in Armstrong's paper is more explicit than the version found in his

Footnote 8 (continued)
I think that the distinct existences argument can be adapted to this additional requirement. The argument can be understood (and I think Armstrong himself saw it this way) as establishing that it is possible to be in φ, but not to believe that one is in φ, *even though the relevant background conditions obtain*. Second: advocates of the constitutive account need not claim that being in φ (plus certain background conditions) *metaphysically* necessitates believing that one is in φ. A much more promising modality in this context is *rational* necessity. Thus, proponents of the constitutive account may concede that the distinct existences argument shows that the connection between φ (plus background conditions) and the belief that one is in φ is not metaphysically necessary. But, they might say, this does not imply that the connection in question is not rationally necessary. So one might well accept the distinct existences argument and yet argue for a constitutive account of self-knowledge. Cf. Stoljar (2018) for a careful and thorough exploration of this way of responding to the distinct existences argument. Stoljar's core idea is that it is one of the requirements of rationality that someone who is in a conscious mental state (and meets other relevant conditions) is rationally required to form the belief that one is in that state. The crucial question, of course, is whether this is actually a requirement of rationality. I have my doubts about this, but I cannot go into it in more detail in this paper.

book. Because Jackson also focuses on the version in Armstrong's paper, I will quote the argument as it appears there:

> "The acquiring of introspective knowledge must consist of the making of (sincere) reports of current mental occurrences, or else a nonverbal apprehension of these occurrences. In both cases the apprehension of the occurrence will have to be *distinct* from the occurrence that is apprehended. But if this is granted, then we can apply Hume's argument about 'distinct existences.' Wherever we have two distinct things, Hume points out, there we can always conceive of the one existing in the absence of the other. It follows that it is logically possible to have a sincere report of a current inner experience, or a nonverbal apprehension of that experience, without the experience existing."[9]

Before assessing this argument, let me try to clarify the content of what may be called *Hume's principle*, that is, the claim that "wherever we have two distinct things, there we can always conceive of the one existing in the absence of the other". As is well known, Hume gained this insight from his thoughts on causation. The inference from the existence of a certain effect to the existence of its cause, he says,

> "is not deriv'd merely from a survey of these particular objects [i.e. the effect and the cause], and from such a penetration into their essences as may discover the dependence of the one upon the other. There is no object, which implies the existence of any other *if we consider these objects in themselves* ..."[10]

I added emphasis to "if we consider these objects in themselves" because I think that this proviso is crucial for an adequate understanding of Hume's principle. Consider the relation between sunburn and sunshine: whenever somebody has sunburn, there must be sunshine, because sunburn *by definition* is caused by sunshine. This example seems to be grist for the anti-Humean's mill: "Hume's principle is wrong," one might say, "look at the sunburn on my neck: although it is certainly distinct from sunshine, my sunburn could not possibly exist without sunshine—since if there was no sunshine, there would not be any sunburn." What is wrong here is that Hume's opponent has considered sunburn not *as it is in itself,* but *as an effect of sunshine*. Of course, if we consider sunburn an effect of sunshine, it could not possibly exist without sunshine. But if we consider sunburn as it is in itself, that is, as *this particular condition of my skin*, it is surely possible that *it* exists without sunshine. Thus, the alleged essential or, as some philosophers prefer to say, internal relation between sunburn and sunshine is ultimately only conceptual: the *concept* "sunburn" may incorporate a specification of the cause of the described phenomenon, but the existence of the phenomenon *itself*, detached from any particular description, does

[9] Armstrong (1963, p. 422). Hume's idea that if two things are distinct, it is possible that one exists without the other can be found in several places throughout the *Treatise*. The idea is most clearly expressed in connection with what Hume calls "perceptions": "Whatever is distinct, is distinguishable; and whatever is distinguishable, is separable by the thought or imagination. All perceptions are distinct. They are, therefore, distinguishable, and separable, and may be conceiv'd as separately existent, and may exist separately, without any contradiction or absurdity" (Hume, 1978, p. 634).

[10] Hume (1978, p. 86, my emphasis).

not necessarily presuppose the existence of a certain cause.[11] Hence, the proviso "if we consider these objects in themselves" is best taken into account by interpreting Hume's principle as a *de re* modal statement, that is, as a statement in which the modal operator is attached to an open sentence:

(Hume's principle) $\forall x \forall y((x \Delta y) \rightarrow \Diamond(\exists w(w = x) \wedge \neg\exists z(z = y)) \wedge \Diamond(\exists w'(w' = y) \wedge \neg\exists z'(z' = x)))$[12]

The next question that has to be considered in order to understand Hume's principle is what the relation of *distinctness* (symbolized by "Δ") is supposed to mean in this context. A very natural candidate for distinctness is *non-identity*. However, as has often been remarked, this reading makes Hume's principle implausibly strong.[13] Consider, for example, the set of individual objects *a*, *b*, and *c*. The set {*a*, *b*, *c*} is not identical to one of its members, say *a*. Still, it is impossible that the set {*a*, *b*, *c*} exists without *a* existing. Similar remarks apply to mereological fusions.

Thus, to keep Hume's principle at least *prima facie* plausible, a stronger reading of "distinctness" is needed. As far as I can see, the best available candidate is a broad notion of mereological distinctness, according to which two objects are mereologically distinct if and only if they share no parts, constituents, or members.[14] In light of this reading of "distinctness," Hume's principle fares much better with respect to apparent counterexamples.

Thus, Armstrong's argument might be reformulated as follows:

I. If two things are such that they share no parts, constituents, or members, then it is possible that the one exists and the other does not. (Hume's principle)
II. My belief that *I currently have a certain conscious experience* shares no parts, constituents, or members with the conscious experience that my belief purports to be about.

[11] I have taken this example from Davidson (1987, pp. 451 f.), who uses it in a different context, however. The example also appears in Hacker (2007, p. 62).

[12] One might object that this reconstruction of Hume's principle, tailored to particulars, does not correspond to the common reading according to which it is a thesis about *properties*. Moreover, one may say that Armstrong needs the property reading for his argument, since mental states are not particulars but properties. Regarding the first point, I refer the reader to Section 7, where the property reading is discussed. As to the second point, I am not sure whether we *must* conceive of mental states as properties — we may as well conceive of them as abstract particulars. Thanks to an anonymous reviewer for pressing me on this point.

[13] Cf. Stoljar (2008) and Wilson (2010). Curiously, Jackson (1973, p. 58) states that he reads "distinctness" as "non-identity". (Actually, Jackson also suggests a second reading—which I will ignore here—according to which "distinctness" means "non-synonymy".) However, Jackson's second counterexample, which I shall refer to in Sect. 4, indicates that he has a stricter reading in mind which coincides with the reading I prefer. See also footnote 25.

[14] Cf. Stoljar (2008, p. 266). Wilson (2010, p. 606) also considers a broad notion of mereological distinctness.

III. Therefore, it is possible that my belief that *I currently have a certain conscious experience* exists, whereas the experience that my belief purports to be about does not, and vice versa.

As can be seen from this reformulation, there are in principle two ways to attack the distinct existences argument: the first way consists in doubting that Hume's principle is true, while the second consists in doubting that the belief that one has a certain experience is distinct from that experience. Since Jackson does not demur at the second premise of Armstrong's argument, but focuses on Hume's principle, I will concentrate on defending the distinct existences argument against attacks on Hume's principle. I will leave the defense of the argument against attacks on its second premise for another occasion.[15]

3 Husbands and wives

Jackson's first counterexample reads thus:

> "A husband is numerically distinct from his wife, but 'I am a husband' entails 'I have a wife' ..."[16]

This counterexample does not suffice for essentially the same reasons that Hume's opponent was wrong with regard to the relation between sunburn and sunshine. The *description* "husband" may indeed entail that the described man is married to a certain woman, but the existence of the man considered *in itself*, detached from any particular description, does not necessarily presuppose the existence of the woman to whom he is in fact married. Even if Prince Philip is in fact married to Queen Elizabeth, it is nevertheless possible that Prince Philip exists and Queen Elizabeth was never born. But let us take a closer look at the issue: I suspect that the crux of the matter is a scope ambiguity with regard to the modal operator, as is so often the case in philosophical disputes.

First, recall that Jackson's claim regarding the entailment relation between "I am a husband" and "I have a wife" does not straightforwardly contradict Hume's principle, because the latter does not concern whether an entailment relation exists between statements, but whether one individual could possibly exist without a different individual. So, let us restate Jackson's claim in such a way that its alleged contradiction with Hume's principle becomes obvious. Continuing with the Prince Philip/Queen Elizabeth example, Jackson's claim could be understood as follows:

[15] The most important contemporary opponents of the second premise are Chalmers (2003), Gertler (2012), and Horgan and Kriegel (2007). All these authors claim that there is a subclass of beliefs about one's phenomenally conscious mental states for which the mental state in question is literally a part of the belief.

[16] Jackson (1973, p. 58). A similar counterexample can be found in Raff (1966, p. 73): "from the nonidentity of *a* and *b* it does not follow that it is possible that *a* exists without *b*, or that *b* exists without *a*. Although President Johnson is not identical with his successor, it is not possible for the successor of Johnson to exist without Johnson existing."

(J) Although Queen Elizabeth's husband does not share any part, constituent, or member with Queen Elizabeth, Queen Elizabeth's husband could not possibly exist without Queen Elizabeth existing.

The rationale behind (J) is that if Queen Elizabeth did not exist, nobody would be married to her. If, in turn, nobody was married to Queen Elizabeth, then nobody would be Queen Elizabeth's husband. To make (J) more transparent, it might be instructive to translate it into logical notation. Let "*a*" stand for Queen Elizabeth, and let "M" be a two-place predicate standing for the relation of *being married to*. Drawing on Russell's symbolism, (J) can be paraphrased as follows:

(J) $((\iota x)\mathrm{M}xa\Delta a) \wedge (\neg\Diamond(\mathrm{E}!(\iota x)\mathrm{M}xa \wedge \neg\mathrm{E}!a))$.

Now, as is well known, sentences such as (J) generally have two readings: on the first reading, the description has a *narrow* scope relative to the modal operator, whereas on the second reading, it takes a *wide* scope. Consequently, on the first reading, (J) turns out to be a *de dicto* modal statement, whereas on the second reading, it turns out to be a *de re* modal statement:

(J*) $\exists x\forall y(\mathrm{M}ya \leftrightarrow y = x) \wedge (x\Delta a) \wedge \neg\Diamond(\exists w(\forall v(\mathrm{M}va \leftrightarrow v = w)) \wedge \neg\exists z(z = a))$;
(J**) $\exists x\forall y(\mathrm{M}ya \leftrightarrow y = x) \wedge (x\Delta a) \wedge \neg\Diamond(\exists w(w = x) \wedge \neg\exists z(z = a))$.

Jackson faces the following dilemma. If we read (J) in the *de dicto* sense of (J*), then it is certainly true, since you can not find a possible world where both statements are true—that is, a world in which someone exists who is married to Queen Elizabeth but in which Queen Elizabeth does not exist. This situation is not possible because nobody can be married to a non-existent person. However, (J*) does not contradict Hume's principle, since, from (J*), the negation of Hume's principle cannot be inferred. The inference is impossible because the quantifiers in the first conjunct of (J*) do not "reach into" the scope of the modal operator.

Now, if we read (J) in the *de re* sense of (J**), it certainly contradicts Hume's principle. However, (J**) is *false,* since even if the actual world is such that there is a person, Prince Philip, who is married to Queen Elizabeth, there certainly is a possible world where that very person, Prince Philip, exists without Queen Elizabeth existing. Of course, in this world, Prince Philip is not married to Queen Elizabeth, and consequently, he lacks the relational property of being married to her (and, presumably, also lacks the property of being a prince). However, the lack of this property does not matter as long as he is cross-world identical to the individual who in the actual world is married to Queen Elizabeth. Being married to Queen Elizabeth is only a *contingent property* of Prince Philip, if you like. In short, if you consider Prince Philip *as he is in himself*—and not *as the husband of Queen Elizabeth*—then it is surely possible that he exists without Elizabeth.

One might admit that Jackson's counterexample fails if it is given that we read Hume's principle in the *de re* sense. However, it could be said that this is merely a Pyrrhic victory for Armstrong. For if we apply the above consideration to mental states, it will become apparent that Armstrong needs the *de dicto* reading of Hume's principle (which is refuted by Jackson's counterexample) as opposed to the *de re* reading (which I am defending). It looks, therefore, as if the proponent of the distinct existences argument faces the following problem: the version of Hume's principle that is most defensible is not the version that Armstrong needs to show what he wants to show, namely that a certain conscious experience can occur without the belief that one is currently having that experience.[17]

To see the problem, recall that according to Armstrong, mental states are defined by their causal roles. Pain, for example, is, roughly speaking, that state of a person—whatever it may turn out to be—that is typically caused by tissue damage and typically causes (1) pain behavior, (2) the belief that one is currently in pain, and (3) the desire to get rid of one's current pain.[18] Only a state that satisfies a certain *description* counts as pain—namely the description "that state of a person which is typically caused by … and typically causes …". In the actual world, according to Armstrong, that description is satisfied by a certain neuronal state, say by C-fibers firing. Thus, the story goes, pain is contingently identical to C-fibers firing.

In a sense, then, the property of playing the pain-role (in particular the property of causing the belief that one is currently in pain) bears a similar relation to C-fibers firing as the property of being married to Queen Elizabeth bears to Prince Philip: it is a *contingent property* of C-fibers firing. Thus there are possible worlds in which C-fibers firing exists without playing the pain-role—which implies that the firing of C-fibers exists without the belief that one is currently in pain. Therefore, it looks as if the desired result would follow quite naturally: it is possible for one to be in pain without the belief that one is in pain. But, wait a minute, something has gone wrong! Recall that according to Armstrong, a state does not count as pain unless it occupies the pain role. In the envisioned possible worlds, however, C-fibers firing does not occupy the pain role. So, in those worlds, C-fibers firing is not identical to pain! Thus, there might be some state contingently identical to pain, namely C-fibers firing, such that it is possible for *it* to occur without the relevant belief. But this is not the desired result. The desired result is that it is possible for *pain* to occur without the relevant belief. It looks, therefore, as if Armstrong would need the *de dicto* reading of Hume's principle to reach the desired conclusion. The *de dicto* reading, however, is refuted by Jackson's counterexample.

The problem may become even clearer if we explain it by analogy with the sunburn example I used above to illustrate the *de re* reading of Hume's principle. According to Armstrong's version of materialism (as I have construed it), pain *by definition* causes a belief to the effect that one is in pain, just like sunburn *by definition is* caused by sunshine. So it seems that just as there is a necessary connection between sunburn and sunshine, there is a necessary connection between pain and the belief that one is in pain. However, as I said in connection with the sunburn

[17] Many thanks to an anonymous reviewer who brought this problem to my attention.

[18] Of course, this way of phrasing the matter is more in line with David Lewis's approach—Armstrong (1968) speaks only of "states apt for bringing about a certain sort of behavior". But I still think Armstrong would approve of my formulation, since his view is broadly in accord with Lewis's approach.

example, we have to look at things in this context as they are in themselves, i.e. independent of any description. Regarding sunburn, this means that we must look at it as *this particular condition of my skin*. And if we do, we will realize that the sunburn (considered as this particular condition of my skin) could exist independently of sunshine. In the same sense, then, we are urged to consider pain for what it is in itself, i.e. independent of the causal role it plays. However, when we do that, we lose sight of the pain, so to speak. Because what we look at then is nothing more than a mere neuronal state. So we are in a quandary: *either* we consider pain *as pain* (and that means, according to Armstrong, that we consider a particular neuronal state as occupying a certain causal role), but in so doing we cannot imagine that pain exists independently of the belief that one is in pain. *Or* we consider pain as it is in itself, independent of any description, but in so doing we likewise cannot imagine that pain exists independently of the belief that one is in pain. For what we then imagine is no longer pain, but merely a particular neuronal state.

I am confident that this problem is not insurmountable though, and there are at least two ways to solve it. The first option is to deny that part of the causal role of pain is to typically cause the belief that one is in pain. However, since this move might smack of a *petitio principii* against philosophers who adhere to the constitutive account of self-knowledge,[19] I prefer the second option: denying that mental states are defined by their causal roles. Rather, "pain" (like other sensation terms) is a rigid designator: across all possible worlds, it picks out that sensation which feels exactly the way pain actually feels, regardless of the causal role it may play in the respective world. Thus, playing the pain role is simply a contingent property of pain. In a world in which some *x* does not play the pain role, *x* does not automatically disqualify as pain—given that *x* feels exactly the way pain actually feels. It is simply not evident at least why the property of causing certain beliefs should be necessary for something to feel exactly the way pain actually feels. This means that when we consider pain as it is in itself, independent of any description, we are not looking at a "bare" neuronal state, but rather at a particular *feeling*. And it is at least intuitively plausible that we can imagine that this particular feeling can exist independently of any belief. Thus, Armstrong would not need the *de dicto* reading of Hume's principle to reach the desired conclusion that it is possible for pain to occur without the relevant belief occurring. All he needs is a semantics of "pain" (and similar terms) according to which it is a rigid designator that picks out a particular feeling across all possible worlds.

Certainly, this suggestion may seem strange at first glance. After all, does it not contradict Armstrong's version of materialism? I have two comments on this: first, as I said, my purpose here is not to defend Armstrong's specific version of materialism, but to consider whether or not the counterexamples commonly advanced against the first premise of the distinct existences argument are convincing. Moreover, the view that "pain" and similar terms are rigid designators does not commit one to a form of anti-materialism à la Kripke.[20] One could hold that sensation terms

[19] See footnote 7.

[20] Kripke (1971).

are rigid and yet still hold to materialism. The crucial step that drives one into the arms of anti-materialism is not the claim that “pain” is a rigid designator, but the claim that “pain = C-fibers firing” is not metaphysically necessary. Nothing I have said commits me to this latter claim. So while my proposal to consider “pain” as a rigid designator referring to a particular feeling does imply a departure from Armstrong’s view that the notion of pain is the notion of a state that occupies a certain causal role, it does *not* imply a departure from materialism.

4 Fathers and daughters

Jackson’s husbands and wives example, then, is hardly likely to put Hume’s principle in distress. Also, doubts about whether the desired conclusion of the distinct existences argument could be reached by means of the *de re* reading of Hume’s principle are dispelled—at least if we are willing to adopt Kripke’s view that “pain” (and similar expressions) are rigid designators. Now, the fact that I brought Kripke into play at this point may bring to mind a new counterexample which seems very well suited for refuting Hume’s principle in its *de re* reading. Consider Kripke’s claim regarding the origin of Queen Elizabeth: she could not have originated from a different sperm and egg from those from which she actually originated.[21] According to Kripke, the existence of Queen Elizabeth necessarily presupposes the existence of the man who was actually her father: King George VI. This does not merely mean that Elizabeth could not have been queen if she had not been George’s daughter—it is not a question of succession to the throne. Also, it is not the trivial conceptual point that “*x* is a child” implies “*x* has parents”. Rather, the point is that there is no possible world in which a given person has different parents than she has in the actual world. According to Kripke, the statement “Elizabeth II could have had different parents than she actually has” is necessarily false. If we think we are imagining a world in which neither George VI nor Elizabeth Bowes-Lyon ever existed and Elizabeth II had other parents, then we are not in fact imagining a world in which Elizabeth II exists, but at most a world in which a person who looks and behaves like Elizabeth II exists.

Of course, Kripke’s claim is not limited to Queen Elizabeth and King George VI, but applies to any person and their parents: it is true for everyone that they could not have existed unless the people who are actually their parents had existed. If we write “*x*P*y*” for “*x* has *y* as biological parent”, the idea can be formulated thus:

$$(1)\quad \forall x\forall y(x\mathrm{P}y\rightarrow\Box(\exists w(w=x)\rightarrow\exists z(z=y))).$$

Moreover, via

[21] Cf. Kripke (1981, pp. 112–113).

(2) $\exists x \exists y (x\mathrm{P}y)$.

and

(3) $\forall x \forall y (x\mathrm{P}y \rightarrow x\Delta y)$.

we arrive at:

(4) $\exists x \exists y ((x\Delta y) \wedge \neg\Diamond(\exists w(w = x) \wedge \neg\exists z(z = y)))$.

It seems, then, that we have a counterexample to Hume's principle here.

My response is that there are good reasons to have doubts about Kripke's origin essentialism since it leads to absurd consequences. Just as the existence of Queen Elizabeth is necessarily tied to the existence of King George VI, the existence of King George VI is necessarily tied to the existence of King George V, whose existence, in turn, necessarily presupposes the existence of Edward VII, and so forth. Because the relation of being necessarily tied to one's ancestors is transitive, it eventually turns out that the existence of Queen Elizabeth is necessarily tied to the existence of some primitive biological organisms at the beginning of evolution.[22] However, this is a very counterintuitive consequence that comes close to being a *reductio*. Thus, one should not base a refutation of Hume's principle on Kripke's origin essentialism.

One might object that the relation of being necessarily tied to one's ancestors is not transitive. Just as the friends of my friends are not perforce my friends, the grandparents of Elizabeth II, though they may necessarily be tied to her parents, are not necessarily tied to Elizabeth II.[23] However, this assumption leads to a contradiction. Assume for the sake of argument that Elizabeth II's grandparents are not necessarily tied to her. This implies that it is possible for Elizabeth II to exist without George V existing:

(5) $\Diamond$((Elizabeth II exists) $\wedge$ $\neg$(George V exists)).

Now, recall that Elizabeth II's existence necessitates the existence of George VI:

(6) $\Box$((Elizabeth II exists) $\rightarrow$ (George VI exists))

Moreover, the existence of George VI necessitates the existence of George V:

(7) $\Box$((George VI exists) $\rightarrow$ (George V exists))

From (6) and (7), it follows

[22] Here, I am heavily indebted to Baumann (2012).

[23] Thanks to an anonymous reviewer who brought this objection to my attention.

(8) $\Box$((Elizabeth II exists) → (George V exists)),

which is equivalent to the negation of (5), to wit:

(9) $\neg\Diamond$((Elizabeth II exists) $\wedge$ $\neg$(George V exists)).

Thus, you cannot have it both ways: claiming that a person's existence is necessarily tied to their parents *and* claiming that that person's existence is not necessarily tied to their grandparents (and more remote ancestors). As soon as you claim that no one could have different parents than they actually have, you are committed to the claim that no one could have different ancestors than they actually have. So Kripke's origin essentialism leads to an inflation of necessary ancestral relations, which in turn leads to the absurd result that it is impossible for anyone to exist without some primitive biological organisms at the beginning of evolution existing. Therefore, we should avoid Kripke's origin essentialism. Without it, however, the example of Elizabeth II and George VI loses its plausibility. So again, Hume's principle is saved.

5 Color and extension: universal reading

Now, let us turn to Jackson's second counterexample:

> "[A]n object's colour is distinct from its extension, but '*A* is coloured' entails '*A* is extended' ..."[24]

Again, this example does not straightforwardly contradict Hume's principle as I formulated it in Sect. 2. To make Jackson's example fit, we must reformulate it along the following lines:

(C1) There are at least two properties—being colored and being extended, for example—such that although they do not share any part, constituent, or member, it is not possible that an object that exemplifies the one but lacks the other exists.[25]

[24] Jackson (1973, p. 58). The example is also presented by Mumford (2004, p. 167) as a counterexample to Hume's principle.

[25] At this point, it becomes clear that Jackson (1973)—contrary to what he himself states—does not understand distinctness merely in terms of non-identity, but has a stronger notion in mind that amounts to what I have called "broad mereological distinctness". The reason for my contention is that *if* Jackson understood distinctness merely in terms of non-identity, he could have invoked a much simpler counterexample alluding to properties expressed by predicates that stand in an analytic entailment relation. For example, he could have said that although *being red* is distinct from *being colored*, "*A* is red" entails "*A* is colored". However, Jackson chooses properties expressed by predicates that do not stand in an analytic entailment relation. Rather, he chooses properties expressed by predicates that are logically independent (and for this reason often cited by proponents of a phenomenological synthetic apriori). The reason that Jackson chooses this example is, I think, that in contrast to the properties of *being red* and *being colored*, it makes good sense to say that the properties of *being colored* and *being extended* are not only non-identical, but also do not share any part, constituent, or member.

The problem with (C1) is that it may be true, but it does not contradict Hume's principle. This becomes obvious when we translate (C1) into logical notation:

$$\text{(C1)}\quad \exists F\exists G(F\Delta G) \wedge [\neg\Diamond(\exists x(Fx \wedge \neg Gx)) \vee \neg\Diamond(\exists x'(Gx' \wedge \neg Fx'))].$$

Note that the negation of Hume's principle reads thus:

$$\text{(Neg)}\quad \exists x\exists y(x\Delta y) \wedge (\neg\Diamond(\exists w(w=x) \wedge \neg\exists z(z=y)) \vee \neg\Diamond(\exists w'(w'=y) \wedge \neg\exists z'(z'=x))).$$

Obviously, (C1) and (Neg) are not equivalent.

One might object that (C1) *does* contradict Hume's principle. To realize this, one might say, we only need to give Hume's principle and (Neg), respectively, a more generous reading according to which the variables range not only over individuals, but also over properties. However, the crucial question then is what formulas such as "$\exists w(w=x)$" are supposed to mean if we assign property values to the variables. It means that the property in question exists, of course. But what does it mean that a property *exists*? My opponent may say that it means that the property in question *is instantiated by at least one individual*. Seen in this light, we might arrive at a reading of (Neg) that would be equivalent to (C1).

However, this is a rather unfortunate move because it confuses the non-existence of a property with its non-instantiation. Note that the non-existence of a property *F* is not the same as the non-instantiation of *F*: *F* may not be instantiated but still exist, whereas, if *F* does not exist, it could not possibly be instantiated. To illustrate the point, compare the property of *both being round and being not-round* and the property of *both being a horse and being winged*. While the first property does not exist (and, consequently, cannot even possibly be instantiated), the second property does exist. Of course, the property of *both being a horse and being winged* is not instantiated, but it is at least possible that it is instantiated.

So, if we do not want to give up the difference between a property's non-existence and its non-instantiation (and I do not think we should do that), then we must admit that (C1) and (Neg) are not equivalent, and that, consequently, Jackson's counterexample poses no threat to Hume's principle.

One might object that it does not seem viable to discard the equivalence of (C1) and (Neg) by invoking the difference between a property's non-existence and its non-instantiation. For Armstrong is firmly opposed to the notion of non-instantiated properties.[26] According to Armstrong's "Principle of Instantiation", for any property to exist it must be instantiated—now, in the past, or in the future—by at least one particular.

Now, we find ourselves in a similar dialectical situation to the one before in connection with the semantics of psychological concepts: it looks as if a departure from Armstrong's own views is necessary to defend Hume's principle. Thus, I argue for the same move I argued for earlier: let us depart from Armstrong's views—for

[26] Cf. Armstrong (1989, pp. 75–82).

my point here, as I have said, is not a defense of Armstrong's overall views, but a defense of the distinct existences argument. Moreover, in the same sense that the view that psychological concepts such as "pain" are rigid designators does not imply a departure from materialism, neither does the recognition of non-instantiated properties impose an anti-naturalistic world view upon us.

To see the point, let us take a brief look at the motives responsible for Armstrong's rejection of non-instantiated properties:

> "Once you have uninstantiated universals you need somewhere special to put them, a "Platonic heaven," as philosophers often say. They are not to be found in the ordinary world of space and time. [...] Such a view is unacceptable to Naturalists, that is, to those who think that the space-time world is all the world that there is."[27]

I think it is relatively easy to dispel the concern Armstrong articulates here. To this end, we need only associate properties with functions that map possible worlds onto sets of world-bound individuals. The property of being red, for example, is associated with the function that, given a possible world as an argument, yields as value the set of red individuals existing in that world. A property exists if and only if it is such that its associated function yields as value a non-empty set of world-bound individuals for at least one possible world (it could but need not be the actual world). In contrast, a property is instantiated if and only if its associated function is such that it yields as value a non-empty set of world-bound individuals at least for the actual world. So, non-instantiated (but existing) properties are those with associated functions that yield as value an empty set for the actual world, but a non-empty set for at least one non-actual possible world. Thus, Armstrong's concern that non-instantiated properties would somehow reside outside of space and time is unfounded. This is because the non-actual possible worlds, for which the functions in question provide non-empty sets of world-bound individuals, are just as spatially and temporally structured as the actual world. Of course, non-actual possible worlds are not actual. But not being actual does not render them realms that are beyond space and time.

So, if Armstrong understands "the ordinary world of space and time" in the sense of "the actual world", he is surely right when he contends that non-instantiated properties "are not to be found in the ordinary world of space and time". But that does not imply that non-instantiated properties somehow reside *beyond* space and time, in a Platonic heaven so to speak. On the contrary, non-instantiated properties *are* in space and time, albeit in the space and time of non-actual worlds. Thus, recognizing non-instantiated properties does not have the dire consequences that Armstrong is concerned about. Certainly, by recognizing non-instantiated properties we violate the wording of Armstrong's theory, but we do not leave the ground of naturalism.

[27] Armstrong (1989, p. 76).

6 Color and extension: trope reading

Still, one might try to defend Jackson's color/extension counterexample from a different angle. One might say that there is a more charitable reading, according to which Jackson is not talking about two distinct *universals*, but two distinct *tropes*: *A*'s being colored at *t* and *A*'s having extension at *t*. Thus, we must understand Jackson along the following lines:

(C2) There are at least two tropes—*A*'s being colored at *t* and *A*'s being extended at *t*, for example—that are such that although they do not share any part, constituent, or member, it is not possible for one to exist without the other existing.

Obviously, (C2) contradicts Hume's principle. But is it true? I am not sure that it is. I wonder, for example, whether *A*'s being colored at *t* and *A*'s being extended at *t* are really such that they do not share a part. At least, my intuition is that the tropes at hand do share a part: the part of *A* which is colored at *t* is the same part of *A* which is extended at *t*. Accordingly, the first conjunct of (C2), and thus (C2) as a whole, would be false.

However, it may be that my intuition is not reliable, so let us abstract from this difficulty. There is, anyway, a more pressing problem. As particulars, tropes must be maximally determinate.[28] Thus, no "generic" color tropes in the sense of being colored but not being of some specific color exist. Similarly, there are no "generic" extension tropes in the sense of being extended but not being of some specific extension.[29] The tropes invoked by Jackson, then, must be of a *specific* color and *specific* extension—for instance, *A's being* red_{45} *at t* and *A's having a surface area of 2.54* cm^2 *at t*. From this perspective, (C2) turns out to be false, since there is certainly a possible world where *A* may be red_{45} at *t,* but does not have a surface area of 2.54 cm^2 at *t,* and vice versa. Hence, it is not impossible for one of the tropes to exist without the other trope existing. Thus, the trope reading of Jackson's second counterexample also fails.

One might object here that Jackson's example can be read differently. It simply says that a color trope cannot exist without an extension trope (or an extension trope without a color trope):

(C3) Although color and extension tropes do not share any part, constituent, or member, it is not possible for a color trope to exist without an extension trope existing, and vice versa.

I agree that the example can be read that way. And I also agree that (C3) is true. But the point is that (C3) does not violate Hume's principle. This may become clear if we formulate (C2) and (C3) in logical notation:

[28] Moltmann (2013, p. 57) speaks of *groundedness* in this context.

[29] I take this point from Gozzano (2008, p. 148).

(C2) $\exists x \exists y(x \text{ is a color trope} \wedge y \text{ is an extension trope} \wedge \neg\Diamond(\exists w(w = x) \wedge \neg\exists z(z = y)) \vee \neg\Diamond(\exists w'(w' = y) \wedge \neg\exists z'(z' = x)))$

(C3) $\neg\Diamond\ (\exists x(x \text{ is a color trope}) \wedge \neg\exists y(y \text{ is an extension trope})) \vee \neg\Diamond(\exists y(y \text{ is an extension trope}) \wedge \neg\exists x(x \text{ is a color trope}))$

So, we are dealing with a similar constellation to that in connection with Jackson's first counterexample. We have two interpretations before us, one of which—the *de re* claim (C2)—contradicts Hume's principle but is false, whereas the other—the *de dicto* claim (C3)—is true but does not contradict Hume's principle. Whichever way you turn it, Jackson's counterexamples don't work.

7 Counterexamples from aposteriori necessity

Against my previous considerations, it could be argued that they miss the point which is actually of interest. So far, I have treated Hume's principle as a principle that is tailored to individual things and cannot easily be applied to universals. At least, that was my strategy in connection with the universal reading of Jackson's example, which invokes the properties of *being colored* and *being extended*. Now, it might be said that this view of Hume's principle is contrary to its reception in contemporary literature—for, today, Hume's principle is usually understood as stating that "there are no necessary connections between distinct *properties*".[30] Thus, it might be said, my considerations may be correct with regard to the individual reading of Hume's principle, but so far nothing has been proved with regard to the universal reading, according to which there are no necessary connections between distinct properties, conceived as universals. Worse still, there are many counterexamples against the universal reading of Hume's principle that I have not yet considered. One type of particularly relevant example concerns aposteriori necessities: though the properties of *being water* and *being hydrogen* do not share any part, constituent, or member, it is necessarily true that, whenever the property of *being water* is instantiated, the property of *being hydrogen* is instantiated as well.[31]

To assess this objection, we must first agree on a formulation of Hume's principle that can be applied to universals. As we have seen, there is the difficulty that my original formulation in Sect. 2 contains formulas such as "$\exists w(w = x)$", whose meaning is not obvious if we assign property values to the variables. However, I have already indicated what the meaning of those formulas is: that it is possible that the property in question is instantiated. Thus, it is natural to define the existence of a property as follows: $\exists w(w = F) \leftrightarrow \Diamond\exists v(Fv)$. Using this definition, Hume's principle can be reformulated as follows:

(Hume's principle*) $\forall F \forall G(F \Delta G) \rightarrow (\Diamond(\Diamond\exists v(Fv) \wedge \neg\Diamond\exists u(Gu)) \wedge \Diamond(\Diamond\exists v'(Gv') \wedge \neg\Diamond\exists u'(Fu')))$.

[30] Stoljar (2008, p. 263, my emphasis).

[31] Cf. McBride (1999, p. 473); Ellis (2001, p. 235).

Translated into natural language (and by harnessing the idea of possible worlds), Hume's principle* amounts to the following:

> If the properties of *being F* and *being G* share no part, constituent, or member,[32] then it is both possible that (1) there is a possible world in which *F* is instantiated, but no possible world in which *G* is instantiated, and that (2) there is a possible world in which *G* is instantiated, but no possible world in which *F* is instantiated.

At first sight, it might not be easy to grasp what is being said here. The iterated modalities in particular cause some difficulties. However, it seems natural to read Hume's principle* as saying that certain constellations of possible worlds are possible, given two distinct properties. Thus, to assess Hume's principle*, we have to consider whether the constellations mentioned, i.e. (1) and (2), are indeed possible or not. Note that, in doing so, our view is not limited to the plurality of possible worlds *as it actually is*, but we can freely vary this plurality at will. As long as the result is logically consistent, anything goes. So in a sense, we are asked to consider counterfactual pluralities of possible worlds. This is something different from merely considering counterfactual possible worlds. We are not considering whether *the actual world* could have been otherwise in this or that respect—rather, we are considering whether *the whole plurality of possible worlds* could have been structured differently. That means that we are not forced to treat the actual world as actual. On the contrary, we are free to consider possible worlds, which are counterfactual from our point of view, as actual, and then see how the space of possibilities unfolds.

To return to the counterexample put forward, let us assume, for the sake of argument, that the properties of *being water* and *being hydrogen* share no part, constituent, or member. The property of *being water*, one might say, just is the property of *being such that it plays the water-role in the actual world*—where *the water-role*, in turn, is cashed out in terms of water's macrophysical features such as *being an odorless, colorless, and tasteless liquid that fills oceans and lakes, quenches thirst,* etc.[33] In light of this, the property of *being hydrogen* is not a constituent of the property of *being water*. Now, let us first consider a plurality of worlds such that there is a possible world in which the property of *being hydrogen* is instantiated, but no possible

[32] One might wonder what it could mean for two properties not to share a part, constituent or member. For it sounds strange, to say the least, to speak of properties having *parts* or *members*. I see the latter in a similar way, but I think that at least the talk of *constituents* of properties makes sense. (For example, the property of being unmarried, although neither a part nor a member, is a constituent of the property of being a bachelor). For the sake of uniformity concerning the meaning of "distinct," however, I retain talk of parts and members even in the context of the property reading of Hume's principle. Those who cannot make sense of the talk of parts or members of properties can at least fall back on the idea of constituents of properties. I am grateful to an anonymous reviewer who encouraged me to clarify this matter.

[33] To avoid possible misunderstandings, let me add that the description "the liquid that plays the water-role in the actual world" is rigid in Kripke's sense: it picks out at every possible world the very liquid which it picks out at the actual world. Thus, given that in the actual world the water-role is played by H_2O, the description picks out H_2O at every possible world. Nevertheless, neither hydrogen nor oxygen is mentioned in the description, not even implicitly.

world in which the property of *being water* is instantiated. This constellation is quite possible. Why? Recall that we are free to consider any counterfactual world as actual. So, let us consider a world as actual in which God, by creating a corresponding natural law, limited the number of electrons in the atomic shell to seven. Thus, the actual world contains hydrogen, but lacks oxygen and, *a fortiori*, water, as a matter of nomic necessity. Now, let us grant, as proponents of so-called *Scientific Essentialism* hold, that nomic necessity *just is* necessity *tout court*: if something is nomically necessary in the actual world, so it obtains in all possible worlds.[34] Thus, relative to the envisaged plurality of worlds, "There is no water" is necessarily true in the same sense as, relative to the actual plurality of worlds, "There is no perpetuum mobile" is necessarily true.

Next, let us consider a plurality of worlds such that there is a possible world in which the property of *being water* is instantiated, but no possible world in which the property of *being hydrogen* is instantiated. To make sense of this possibility, recall that, *ex hypothesi*, the property of *being water* just is the property of *being such that it plays the water-role in the actual world*. Now, it is certainly not logically inconsistent to assume that, when God created the actual world, He created a world in which not H_2O, but XYZ played the water-role. Thus, relative to the plurality of worlds envisaged, "water = XYZ" is necessarily true. Assume, furthermore, that God, when He created the actual world, created a natural law according to which an element such as hydrogen is nomically impossible. Again, add the assumption that what is nomically impossible is impossible *tout court,* and you arrive at the result that the plurality of worlds envisaged is quite possible. Thus, the counterexample from aposteriori necessity fails. Aposteriori necessities are compatible with Hume's principle*.

Let me briefly discuss two objections that might be raised. The first is that the property of *being water* is not identical to the property of *being such that it plays the water-role in the actual world*. Rather, it is identical to the property of *being H_2O*. Thus, it simply would not follow from the fact that XYZ plays the water-role in the actual world that "water = XYZ" is necessarily true. That's all well and good. But then the properties of *being water* and *being hydrogen* would no longer be distinct in the relevant sense of sharing no part, constituent, or member. Recall that Hume's principle* is not disproved until one finds a case in which the following holds:

> There are two properties of *being F* and *being G* such that they share no part, constituent, or member, and it is neither possible that (1) there is a possible world in which *F* is instantiated, but no possible world in which *G* is instantiated, nor that (2) there is a possible world in which *G* is instantiated, but no possible world in which *F* is instantiated.

If the property of *being water* is identical to the property of *being H_2O*, then the first conjunct of the above requirement ("there are two properties of *being F* and *being G* such that they share no part, constituent, or member") is not satisfied—for the property of *being H_2O* contains the property of *being hydrogen* as a constituent.

[34] Cf. Ellis (2001, p. 253).

Thus, the assumption with which I started my consideration—namely, that the property of *being water* is identical to the property of *being such that it plays the water-role in the actual world*—was an assumption I made *in favor of* my opponent. For only in this way does it remain guaranteed that the property of *being water* is distinct from the property of *being hydrogen*. However, if we give up this assumption and instead adopt the view that the property of *being water* is identical to the property of *being H_2O*, then the distinctness between the property of *being water* and the property of *being hydrogen* is lost. And this would be disadvantageous for my opponent.

The second objection is that we shouldn't simply assume that nomic necessities are necessities *tout court*. *Scientific Essentialism* is just wrong. Thus, from the fact that there are natural laws preventing the formation of water in the actual world, it simply does not follow that "There is no water" is necessarily true. Again, that's all well and good. However, if we deny *Scientific Essentialism* and insist that natural laws are just contingently true, then the motive to attack Hume's principle is lost. The motive was precisely to show that Hume was wrong in his denial of necessary relations in nature. Someone who denies the necessity of natural laws should not be attacking, but embracing Hume's principle.

So again, the assumption I made—assuming the truth of *Scientific Essentialism*—was made *for the benefit* of my opponent. If we give up this assumption and assume that the laws of nature which are valid in the actual world are not valid in all possible worlds, then it becomes extremely difficult, if not impossible, for my opponent to find examples of distinct properties which must be exemplified together.

8 Summary

Jackson's argument against Armstrong's distinct existences argument fails across the board. The examples he comes up with to disprove Hume's principle are ineffective: the first suffers from an ambiguity regarding the scope of definite descriptions, whereas the second oscillates between a universal reading and a trope reading. Moreover, the other counterexamples against Hume's principle fail as well: they either commit us to at least *prima facie* implausible doctrines, such as Kripke's origin essentialism, or—as in the case of aposteriori necessities—they are not in conflict with Hume's principle.

However, we have also seen that the defense of the distinct existences argument comes at a certain cost for Armstrong. Not only have we been forced to abandon his thesis that mental states are defined in terms of their causal roles, but also his thesis that there are no non-instantiated properties. As unfavorable as this result may be for Armstrong—for defenders of the distinct existences argument it is rather advantageous. For it makes clear that the argument does not depend on specific (and quite controversial) assumptions from other sections of Armstrong's philosophy, but can be sustained independently of Armstrong's specific philosophical views.[35]

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Anstey, P. R., & Braddon-Mitchell, D. (eds.) (forthcoming), *A materialist theory of the mind: 50 years on*.

Armstrong, D. M. (1963). Is introspective knowledge incorrigible? *Philosophical Review, 72*, 417–432

Armstrong, D. M. (1968). *A materialist theory of the mind*. Routledge.

Armstrong, D. M. (1976). Incorrigibility, materialism and causation. *Philosophical Studies, 30*, 125–127

Armstrong, D. M. (1989). *Universals—An opinionated introduction*. Westview Press.

Baumann, P. (2012). On the inflation of necessities. *Metaphysica, 13*, 51–54

Bilgrami, A. (2006). *Self-knowledge and resentment*. Harvard University Press.

Boyle, M. (2011). Transparent self-knowledge. *Proceedings of the Aristotelian Society Supplementary, 85*, 223–241

Chalmers, D. (2003). The content and epistemology of phenomenal belief. In Q. Smith & A. Jokic (Eds.), *Consciousness: New philosophical perspectives.* (pp. 220–272). Oxford University Press.

Coliva, A. (2009). Self-knowledge and commitments. *Synthese, 171*, 365–375

Davidson, D. (1987). Knowing one's own mind. *Proceedings and Addresses of the American Philosophical Association, 60*, 441–458.

Ellis, B. (1976). Avowals are more corrigible than you think. *Australasian Journal of Philosophy, 54*, 116–122

Ellis, B. (2001). *Scientific essentialism*. Cambridge University Press.

Gallois, A. (1996). *The world without the mind within*. Cambridge University Press.

Gertler, B. (2012). Renewed acquaintance. In D. Smithies & D. Stoljar (Eds.), *Introspection and consciousness.* (pp. 93–127). Oxford University Press.

Gozzano, S. (2008). Trope's simplicity and mental causation. In S. Gozzano & F. Orilia (Eds.), *Tropes, universals and the philosophy of mind.* (pp. 133–154). Ontos.

Hacker, P. M. S. (2007). *Human nature: The categorial framework*. Blackwell.

Heal, J. (2002). On first-person authority. *Proceedings of the Aristotelian Society, 102*, 1–19

Horgan, T., & Kriegel, U. (2007). Phenomenal epistemology: What is consciousness that we may know it so well? *Philosophical Issues, 17*, 123–144

Hume, D. (1978). *A treatise of human nature*. Oxford University Press.

Jackson, F. (1973). Is there a good argument against the incorrigibility thesis? *Australasian Journal of Philosophy, 51*, 51–62

[35] For valuable suggestions and helpful comments on earlier versions of this paper, I would like to thank three anonymous reviewers for this journal, the participants of Thomas Spitzley's graduate seminar at the University of Duisburg-Essen, and the audience of a lecture I delivered at Goethe-University Frankfurt.

Kripke, S. (1971). Identity and necessity. In M. K. Munitz (Ed.), *Identity and individuation.* (pp. 135–164). New York University Press.

Kripke, S. (1981). *Naming and necessity*. Blackwell.

Lewis, D. (1972). Psychophysical and theoretical identifications. *Australasian Journal of Philosophy, 50*, 249–258.

McBride, F. (1999). Could Armstrong have been a universal? *Mind, 108*, 471–501

Moltmann, F. (2013). *Abstract objects and the semantics of natural language*. Oxford University Press.

Mumford, S. (2004). *Laws in nature*. Routledge.

Pappas, G. S. (1976). Incorrigibility and central-state materialism. *Philosophical Studies, 29*, 445–456

Raff, C. (1966). Introspection and incorrigibility. *Philosophy and Phenomenological Research, 27*, 69–73

Shoemaker, S. (1994). Self-knowledge and inner sense. *Philosophy and Phenomenological Research, 54*, 249–314

Stoljar, D. (2008). Distinctions in distinction. In J. Kallestrup & J. Hohwy (Eds.), *Being reduced: New essays on causation and explanation in the special sciences.* (pp. 263–279). Oxford University Press.

Stoljar, D. (2018). Introspection and necessity. *Noûs, 52*, 389–410

Verges, F. G. (1974). Jackson on incorrigibility. *Australasian Journal of Philosophy, 53*, 243–250

Wilson, J. (2010). What Is Hume's Dictum, and why believe it? *Philosophy and Phenomenological Research, 80*, 595–637

Zimmerman, A. Z. (2006). Basic self-knowledge: answering peacocke's criticisms of constitutivism. *Philosophical Studies, 128*, 337–379