Philosophy of Science (2022), **89**, 661–682
doi:10.1017/psa.2022.7

PSA PHILOSOPHY OF SCIENCE ASSOCIATION
PHILOSOPHY OF SCIENCE

ARTICLE

Why Mental Disorders are not Like Software Bugs

Harriet Fagerberg

Department of Philosophy, King's College London, London, UK and Institut für Philosophie, Humboldt-Universität zu Berlin, Berlin, Deutschland
Email: harriet.fagerberg@kcl.ac.uk

(Received 28 July 2020; revised 29 March 2021; accepted 20 October 2021; first published online 11 February 2022)

Abstract

According to the Argument for Autonomous Mental Disorder (AAMD), mental disorder can occur in the absence of brain disorder, just as software problems can occur in the absence of hardware problems in a computer. This article argues that the AAMD is unsound. I begin by introducing the "natural dysfunction analysis" of disorder, before outlining the AAMD. I then analyze the necessary conditions for realizer autonomous dysfunction. Building on this, I show that software functions disassociate from hardware functions in a way that mental functions do not disassociate from brain functions. It follows that mental disorders are brain disorders necessarily.

1. Introduction

According to the Argument for Autonomous Mental Disorder (AAMD), mental disorder can occur in the absence of brain disorder, just as software problems can occur in the absence of hardware problems in a computer. This article argues that this argument is unsound and should be rejected.

The AAMD serves two primary philosophical purposes. Firstly, it is employed to counter the antipsychiatric contention that mental disorders that are not brain disorders are not *real* (Papineau 1994; Kingma 2013; cf. Szasz 1960). Secondly, it is invoked to show that the blanket doctrine that *all* mental disorders are ipso facto brain disorders fails to hold up (Wakefield 2014a; cf. Insel et al. 2010).[1] According to its proponents, the argument from the computer analogy establishes that real, scientifically respectable mental disorder can occur in the absence of brain dysfunction, and that this is compatible with physicalism and with our best philosophical theories of disorder.

[1] This kind of view features prominently in the scientific and biomedical discourse. It was explicitly adopted by the National Institute of Mental Health's Research Domain Criteria (RDoC) initiative at the beginning of the last decade—a move that continues to cause controversy (Insel et al. 2010; Insel and Cuthbert 2015; cf. Borsboom et al. 2019).

As I shall proceed to show, the AAMD is unsound. There is a crucial disanalogy between software-hardware and mind-brain. Not all software functions are hardware functions, but *all* mental functions are brain functions. As such, the analogy fails, and the argument fails to support its stated conclusion.

I begin by introducing the natural dysfunction analysis of disorder, before outlining the AAMD as per Wakefield and Papineau. I then analyze the general conditions under which a failure to perform some process or effect E constitutes a dysfunction of an item X. Building on this, I explain why some computer software functions fail to satisfy these conditions in respect of the hardware. Because some software functions are not selected effects of the hardware, autonomous software dysfunction is possible. In contrast, *all* mental functions are necessarily selected effects of the brain. This is the crucial disanalogy between software-hardware and mind-brain. The AAMD thus fails, and its conclusion must be rejected. Properly understood, autonomous mental dysfunction cannot obtain.

2. The natural dysfunction analysis

I shall first consider, in general terms, the theory of medical disorder from which the argument for autonomous mental disorders proceeds.[2] We will call this view the "natural dysfunction analysis."

According to the natural dysfunction analysis, disorder is natural dysfunction, and natural function should be construed in accordance with the etiological or evolutionary theory of biological function.

In other words, something like the following is assumed:

> It is a *natural function* of an item X in an organism O to do that "which items of X's type did to contribute to the inclusive fitness of O's ancestors, and which caused the genotype, of which X is the phenotypic expression, to be selected by natural selection" (Neander 1991, 174).
>
> If some item X is unable to (adequately) perform one of X's natural functions, then that constitutes a *natural dysfunction* of X.
>
> *Medical disorder* is natural dysfunction.

The formulation "item," in the preceding text, is deliberately noncommittal as to whether the locus of disorder is the mechanism, system, organ, or something else, but paradigmatically, it will be an organ or a part of an organ.

To give an indication of how this analysis can be applied in practice, one might reason that the inability of some particular human heart to pump blood constitutes a "medical disorder," because "blood pumping" is an effect that contributed to the inclusive fitness of our ancestors, such that the genotype that codes for the development of the phenotype "heart" was naturally selected.

Both Papineau's and Wakefield's versions of the AAMD proceed from views akin to the natural dysfunction analysis. Wakefield famously defends what is often referred to

[2] I am using the term "medical disorder" or "disorder" in a broad, general sense to refer to phenomena legitimately inherent to the medical realm, including psychopathology (roughly in accordance with Wakefield's usage; see Wakefield [2014b]).

https://doi.org/10.1017/psa.2022.7 Published online by Cambridge University Press

as the "harmful dysfunction analysis" of medical disorder. On this view, medical disorder is jointly composed of (1) a value relative harm component—"some harm or deprivation of benefit to the person" (Wakefield 1992, 384)—and (2) a value-neutral dysfunction component—"the inability of some internal mechanism to perform its natural function" (384). Natural functions, according to Wakefield, are effects that are "part of the evolutionary explanation of the existence and structure" (Wakefield 1992, 384) of a mechanism. Similarly Papineau, following Neander (1983), takes disorder to be constituted by "biological dysfunction," which he defines as: "items not producing the effects . . . in virtue of which they were naturally selected" (Papineau 1994, 81). Papineau also invokes an additional, seemingly evaluative, condition. According to Papineau, a biological dysfunction "only counts as an illness if it is also in some sense incapacitating" (Papineau 1994, 81).

For the purposes of the present argument, I shall proceed as if dysfunction were both necessary and sufficient for disorder. Whether a value-relative criterion is needed is an interesting question, but one that would take us too far afield. See Cooper (2017) and Wakefield (2014b) for recent discussion of key issues.

3. The argument for autonomous mental disorder

Different versions of the AAMD share the contention that mental disorders can occur in the absence of any disorder of the brain, yet still be *real* disorders with a scientifically respectable physical basis—just as software problems can occur in the absence of hardware problems in a computer (Boorse 1976; Papineau 1994; Arpaly 2005; Wakefield 2006; Cooper 2007; Graham 2013; Kingma 2013; Jefferson 2020). In the main body of this article, I consider a version of the argument that explicitly proceeds from the natural dysfunction analysis, following Wakefield and Papineau. I consider the implications for nonnaturalists about mental disorder in the final section of this article (8.3).

3.1. Wakefield

According to Wakefield, there are two scenarios in which a condition counts as a mental disorder but not a brain disorder.

Firstly, a mental disorder is autonomously mental when a mental dysfunction is realized by a "normal" nondysfunctional brain state. The possibility of this is implied by the possibility of software problems in the absence of hardware problems. Wakefield invites us to consider the computer analogy:

> A computer's software runs in the hardware and therefore a given state of the software while running is always at any given moment identical to some hardware state, but the software can nevertheless malfunction even though there is no malfunction whatever in the hardware. It is true that every software malfunction has some hardware description; that is not at issue. Rather, the point is that a software malfunction need not be a physical hardware malfunction. Analogously, even if all mental states are physical states, it does not follow that a mental dysfunction is a physical dysfunction. (Wakefield 2006, 129)

He repeats this contention in a 2014 critique of the RDoC's commitment to the view that mental disorders are necessarily brain disorders (Wakefield 2014a), and again in two recent (Wakefield 2017, 2020) papers on addiction:

https://doi.org/10.1017/psa.2022.7 Published online by Cambridge University Press

> The invalidity of "all mental disorders are in the brain, therefore all mental disorders are brain disorders" is suggested by the invalidity of the analogous argument: All computer software runs in computer hardware, therefore all software malfunctions must be hardware malfunctions. (Wakefield 2017, 57)

Secondly, Wakefield takes it that a mental disorder is autonomous when it is multiply realized neurobiologically. He does not offer an independent argument for why the multiple realizability of a mental dysfunction would negate its being realized by a brain dysfunction—he merely cites Brülde and Radovic (2006), who in turn cite Svennson (1990). More recently, Jefferson (2020) has offered an argument with a similar implication. I agree with Wakefield that multiply realized mental dysfunctions (i.e., the second scenario) raise different issues from the first (i.e., the analogy to software-hardware). These are separate contexts, which regrettably are often confused and conflated in the literature. However, I reject Wakefield's contention that whether a mental dysfunction is multiply realized is of relevance to whether it is, or is not, a brain dysfunction. This simply does not follow, and to think otherwise is a deductive error. I shall outline an argument for this in section 8.2.

For now, we shall leave multiple realization to one side and focus on the first case: the argument for autonomous mental disorder from the computer analogy.

3.2. Papineau

Papineau's version of the AAMD remains the most explicit and persuasive. In what follows, I lay out the argument as originally offered by Papineau. I will refer back to Papineau in assessing the soundness of the AAMD (section 4 onward). Somewhat reconstructed, Papineau's argument can be analyzed in seven steps:

1. Disorder just *is* natural dysfunction.

Papineau holds that disorder is biological dysfunction, and that biological dysfunction is a matter of "items" not doing whatever it is they were naturally designed to do. This, as explained in section 2, is a version of the natural dysfunction analysis.

2. Software dysfunction need not imply hardware dysfunction.

Like Wakefield, Papineau appeals to the computer metaphor:

> Suppose that you and I are both using MS Word 5.0 as our word processing program, but that you are working on a PC while I am working on a Macintosh. Now suppose that there is some bug in the program. For example, suppose that whenever either of us tries to double-space a highlighted section, that section gets deleted. This obviously wouldn't show that there was anything physically wrong with our machines . . . . The logic circuits are all working as they are supposed to. Rather, the fault lies entirely at the software level. (Papineau 1994, 79)

https://doi.org/10.1017/psa.2022.7 Published online by Cambridge University Press

The hardware is doing everything it is supposed to be doing—in terms of responding correctly to the software code—and yet something is clearly going wrong. That "going wrong" cannot however be attributed to the hardware of either computer. Indeed, the malfunction, in this case, is unique to the software.[3] Papineau thus establishes the possibility of "autonomous software dysfunction." In the case of computer processing, dysfunctions of the supervenient property (the software) can occur in the absence of any dysfunction whatsoever of the realizing property (the hardware).

3. Mind-brain is analogous to software-hardware in all relevant respects.

An analogy is not in and of itself sufficient for an argument; it also needs to be an apt one. More specifically, the analogy has to apply in all ways relevant to the case.[4] Papineau seems to take the applicability of the analogy to follow from nonreductive physicalism. The thought is perhaps something like this: Because the computer analogy is frequently invoked to illustrate the phenomenon of multiple realization, and because mental types are generally multiply realized in distinct neurobiological types, the computer analogy applies to the case of mind-brain.

As I shall show, the computer analogy does not in fact apply to the case of mind-brain (and, in any case, it would not follow from nonreductive physicalism that it does). In other words, I will reject premise 3.

4. From 2 and 3, mental dysfunction need not imply brain dysfunction.

Number 4 is the logical consequence of 2 and 3. If it is true that software dysfunction can occur in the absence of hardware dysfunction (2), and it is true that mind-brain mirrors software-hardware in this respect (3), then mental dysfunction need not imply brain dysfunction (4).

5. Mental dysfunction can be "natural."

On Papineau's view, it is not just paradigm biological items, like organs and limbs, that can have "natural functions." Mental "items" (systems, mechanisms, modules, or the like) are also selected "in the course of genetic evolution and individual learning" (1994, 81) for the production of certain mental effects. If natural mental functions are possible, then natural mental *dys*functions—where a mental item fails adequately to yield one of its selected effects—are presumably also possible.

6. From 4 and 5, natural mental dysfunction need not imply natural brain dysfunction.

[3] One might wonder whether, in Millikanian terms, the hardware has the *derived* proper function of performing specific software functions such as double spacing (Millikan 1989). If so, then "failure to double space" could count as a (derived) hardware malfunction. It would appear to be an assumption of Papineau's argument that this is not the case. For present purposes, we shall put this issue to one side, and grant Papineau's analysis in this regard.

[4] Roughly speaking, a fact is "relevant" if its truth/falsity would make a difference to whether the analogy does (or does not) imply autonomous mental disorder.

https://doi.org/10.1017/psa.2022.7 Published online by Cambridge University Press

If mental dysfunction need not imply brain dysfunction (4), and mental dysfunctions can be "natural" (in the sense of being a failure in a naturally selected effect, as per 5), then natural mental dysfunction need not imply natural brain dysfunction.

7. From 6 and 1, autonomous mental disorder is possible.

Because natural mental dysfunction need not imply natural brain dysfunction (per 6), and because disorder is natural dysfunction (per 1), mental disorder need not imply brain disorder. Mental disorders that occur in the absence of brain disorder are thus "autonomous."

3.3. The virtues of autonomy

We are thus led to conclude the following. The mind is nothing over and above the brain, yet mental disorder does not necessarily imply brain disorder. The former claim does not infringe upon the latter; indeed the "autonomy" of mental disorders follows from standard run-of-the-mill nonreductive physicalism about the mind-brain relation. Nevertheless, mental disorders are still *real* medical disorders, constituted by natural dysfunction. The AAMD thus dissolves a prima facie puzzle as to how nonreductionism, physicalism, and realism can be combined in psychiatry.

In philosophy of psychiatry, being able to combine physicalism, realism, and nonreductionism about mental disorder is attractive in that their theoretical antonyms (reductionism, nonphysicalism, and eliminativism about mental disorder) are generally viewed as undesirable. It is not obvious prima facie, however, how nonreductionism (mental disorders are *not* brain disorders), physicalism (mental disorders are physical), and realism (mental disorders are real) can be combined. At first sight, if we accept that everything that exists is physical (physicalism), and that mental disorders are real (realism), then we are led to conclude that mental disorders are *really* physical disorders (reductionism). We could instead reject physicalism—and maintain that mental disorders are real but nonphysical—but this would be intolerable to most contemporary philosophers and scientists. If we maintain, however, that mental disorders are *not* physical disorders (nonreductionism), but accept that everything that exists is physical, then it seems we are led to conclude that mental disorders do not *really* exist (eliminativism).

This latter inference is made in some antipsychiatric contexts (prominently by Szasz 1960) and was the target of Papineau's original argument (1994). Thus, the AAMD provides an escape route from an underlying conflict in theoretical desiderata that has plagued psychiatry and the philosophy thereof for decades. This ideologically convenient property of the AAMD may go some way to account for the relative lack of critical resistance to this line of argument in the literature.

4. What is it to be a dysfunction of X?

Having outlined the natural dysfunction analysis and how it has been combined with the software-hardware analogy to argue for the autonomy of mental dysfunction, some further clarifications are in order.

Let us first specify the position to which proponents of the AAMD must commit. Proponents must commit to the view that, just as software dysfunction does not entail

https://doi.org/10.1017/psa.2022.7 Published online by Cambridge University Press

hardware dysfunction, natural mental dysfunction does not entail natural brain dysfunction. In other words, the following can obtain:

> A particular subject at a particular time could be instantiating a natural mental dysfunction in the absence of natural brain dysfunction.

In the case that natural mental dysfunctions either: (a) are necessarily natural brain dysfunctions or (b) necessarily co-occur with natural brain dysfunctions, the preceding cannot obtain. My view is that (a) is true. Natural mental dysfunctions cannot be instantiated in the absence of natural brain dysfunctions because natural mental dysfunctions *are* natural brain dysfunctions.

To understand this position, we must understand what it means for a mental dysfunction to *be* a brain dysfunction. More generally, we need to know what conditions some particular state would need to satisfy in relation to an item to count as a dysfunction of that item. In other words, what does it take for the failure of some process E to constitute a dysfunction of some item X?[5]

In what follows, let "X" be a schematic letter denoting some biological trait or item that may have one or more natural functions. X will, paradigmatically, be an organ, such as the heart. But X might be a mechanism, such as the urea cycle (a metabolic pathway). Let "E" denote some process or effect. For example, E might be the process of pumping blood, or the process of producing urea from ammonia. We will use "¬E" to denote the complement of E, the property of not doing E: the state of not pumping blood or the state of failing to produce urea from ammonia.[6]

There are two conditions that need to be met for X's being ¬E to constitute a dysfunction of X:

1) X instantiates ¬E
2) E is a natural function of X

(1) Reminds us that it has to be X, and not something else, that fails to E, for failing to E to be a dysfunction of X. (2) tells us that ¬E is a dysfunction of X only if E is a (natural) function of X. X may be ¬E, but unless X has E as one of its functions, then this is not a dysfunction of X. However vigorously and purposefully I flap my arms, I am unable to fly, yet this sad state of affairs does not constitute a dysfunction because "flight" is a not a natural function of my arms, nor of any other part of my body. Accordingly, my inability to fly cannot logically constitute a natural dysfunction of any part of me.

We can now apply this to the brain. For the failure or some mental process ME (¬ME) to be a dysfunction of the brain, B, it has to be the case that:

[5] Note that I am using "dysfunction of X" and "X dysfunction" interchangeably to mean the same thing as well as, for example, "dysfunction of the hardware" and "hardware dysfunction," etc.

[6] ¬E technically has to be read as including the full range of not doing E *properly*, including doing E too much. It also has to be read as excluding instances where X cannot perform E *only* due to the fact that X is lacking one of the environmental preconditions for normal functioning (e.g., an appliance that cannot perform its function due to not being connected to electricity; see Garson [2019]). Thanks to an anonymous reviewer for raising this issue.

https://doi.org/10.1017/psa.2022.7 Published online by Cambridge University Press

1) B instantiates ¬ME
2) ME is a natural function of B

We can conclude now that the AAMD entails a commitment to the view that there is logical "elbow room" for real, medical mental disorders that fail to satisfy conditions 1 and 2. In other words, according to the AAMD, for some mental effects, ME, the failure of that effect, ¬ME, can be a natural dysfunction and hence a *real* mental disorder, without satisfying both 1 and 2 in relation to the brain. As noted, I do not think this can obtain. To see why, we need first to get clear on *how* software dysfunction can occur in the absence of hardware dysfunction in a computer.

5. How is autonomous software dysfunction possible?

To understand what it would mean for a mental dysfunction to be "autonomous" from brain dysfunction in the sense suggested by the analogy to software-hardware, we need first to understand precisely *how* autonomous software dysfunction obtains in the case of classical computers.[7]

In what follows I shall show how autonomous software dysfunction (software dysfunction in the absence of hardware dysfunction) is possible. I will then abstract to a more general model, specifying the conditions which "supervenient" dysfunctions in general (i.e., realizer-independent dysfunctions such as autonomous software dysfunction or autonomous mental dysfunction) have to satisfy. I will call this set of conditions the "Autonomous Model." I shall go on to argue that the Autonomous Model is true description of a state that *can* obtain in the case of software and hardware, but that *cannot* obtain in the case of mind and brain, and that the AAMD thus fails.

First, let us return to the software-hardware analogy. Let HW be the hardware, SW be the software, and SE be some software process, such as Papineau's "double-spacing." The following can obtain:

1) HW instantiates ¬SE

But not:

2) SE is a function of HW

In the in the preceding text, the hardware instantiates "failure to double-space" in that failure to double-space is instantiated *in* the hardware. However, the hardware

[7] What I and others in this area refer to as "autonomous software dysfunction" or "software dysfunction without hardware dysfunction" is sometimes called "programming error" or "software design error" in the literature on computation. There is some controversy within this literature as to whether software design errors count as genuine instances of miscomputation, or whether this status should be reserved for so-called operational malfunctions—that is, breakdowns of the actual internal operations of the computational system (Fresco and Primiero 2013; Dewhurst 2014). This is a complicated issue, and I am largely agnostic as to whether or not "design error" should count as "miscomputation." However, in the sense that there is a failure to yield the effects intended by the software designer, "software design errors" do count as dysfunctions (for discussion, see Tucker 2018; Coelho Mollo 2021).

https://doi.org/10.1017/psa.2022.7 Published online by Cambridge University Press

does not have "double-spacing" as a function. Thus, with regard to the hardware, SE satisfies 1 but not 2. Accordingly, per the conditions outlined, ¬SE simply cannot be a dysfunction of HW. Thus far, this case is the same as me flapping my arms and failing to fly. If this were the end of the story, there simply would be no dysfunction—whether of the hardware or of the software. However, this simple picture is complicated by the fact that SE is a function of a supervenient property of the hardware (i.e., the software, SW):

1) HW instantiates ¬SE

And:

3) SE is a function of SW

But not:

2) SE is a function of HW

Thus, when ¬SE (failure to double-space) is instantiated in the hardware, a dysfunction *is* occurring, and that dysfunction is located *in* the hardware. However, that dysfunction is not a dysfunction *of* the hardware.[8]

Because this situation can obtain, autonomous software dysfunction is possible. These relationships can be expressed in more general terms as follows. X is an item (organ, mechanism etc.), E is a process, and S is a property tthat supervenes on X. Call this the Autonomous Model.

1) X instantiates ¬E

And:

3) E is a function of S

But not:

2) E is a function of X

The Autonomous Model describes the components and relations necessary for autonomous software dysfunction and "supervenient" dysfunctions more generally to be possible—stated simply, 1 and 3, but not 2. The reason why "failure to double space" is a dysfunction of the software and not of the hardware is that "double-spacing" is a function of the software and *not* a function of the hardware.

It now becomes clear that conforming to the Autonomous Model depends on some functions of the supervenient property, the software, *not* being functions of the realizing property, the hardware. (Otherwise, ¬E would be instantiated by X in addition to

[8] Some readers will be familiar with the AAMD as advanced by Graham (2013). I take the preceding to clarify the "in/of" distinction that Graham postulates in his version of the argument for autonomous mental disorder from the computer analogy.

https://doi.org/10.1017/psa.2022.7 Published online by Cambridge University Press

E being a function of X, and E would as such satisfy conditions 1 and 2, as outlined, and accordingly ¬E would be an X dysfunction.) So to some extent, the functions of S need to "come apart" from the functions of X.

Let us call this "functional separability." Autonomous software dysfunction is possible because functional separability obtains in the case of hardware-software. The question from here is: Does functional separability obtain in the case of mind-brain (such that mind-brain might conform to the Autonomous Model and, in turn, autonomous mental dysfunction might conceivably occur)? To answer this question, we need to know *why* functional separability obtains in software and hardware.

6. Why is there "functional separability" in hardware-software?

Why is it that *some* software functions are not also hardware functions? What accounts for the functional separability of software-hardware (and thus the possibility of autonomous software dysfunction)?

The etiological theory of function (as outlined in section 2) tells us that E is a function of X when E is a selected effect of X. More precisely, E is a function of X if and only if E is an effect that was causally efficacious in the natural selection and design of X through evolutionary history. The natural selection of X through evolutionary history is X's *functional etiology*. According to the etiological theory, functions depend upon and are the outcome of these histories of "design" or selection.

Because hardware software are artifacts (and as such not products of natural selection) they lack natural functions in a strict sense. However, artifacts too have etiological functions, which depend upon their histories of intentional artifact design. We might then say that E is the function of artifact X if and only if E is an effect that was causally efficacious in the intentional design of X through X's history of design and selection (i.e., through X's functional etiology). Now, are all software processes effects that were causally efficacious throughout X's history of design? In other words, are software processes necessarily selected effects of the hardware?

Indeed, they are not. Classical computers are general purpose processing machines.[9] They are designed to run software, not some particular kind of software. The effect of the hardware, which was causally efficacious in its design and thus determines its etiological function, is its capacity to respond accurately and predictably to software code. The particular functions of some particular piece of software (such as the capacity to double-space text within a word processing program) need not have played any role whatsoever in the design of the hardware.

Software and hardware come about using separate functional etiologies that are sensitive to distinct selection pressures. Indeed, one might imagine that the hardware designers were totally unaware that the hardware they were designing would eventually come to run word processing software—never mind one with a function such as "double-spacing." The capacity to double-space, in such a scenario, is entirely causally impotent in the etiology of the hardware. The effect that explains the structure and existence of the hardware is its general-purpose processing capacity, *not* any

[9] There are computational systems (sometimes called "dedicated computers") where the software in so embedded in the hardware that the hardware really is designed for a single hardwired, computational process and there is no "software" in the modern sense. These, however, are not computers in which autonomous software dysfunction would be possible.

https://doi.org/10.1017/psa.2022.7 Published online by Cambridge University Press

particular software or the peculiar software-functions thereof. It follows straightforwardly from the etiological theory of function that "double-spacing" is *not* a hardware function.

However, "double-spacing" *was* causally efficacious in the development of a different product—the software. One might imagine that the software designers meditated deeply over how to best configure processes such as "double-spacing" into the code for their word processing software, whilst the hardware designers enjoyed their blissful ignorance of these software-specific concerns. This is what gives rise to functional separability. Because "double-spacing" and other particular software effects are selected effects of the software but *not* of the hardware they logically cannot constitute hardware dysfunctions. Thus, the possibility of autonomous software dysfunction arises.

We might update the Autonomous Model accordingly:

1) X instantiates ¬E

And:

3) E is a selected effect of S

But not:

2) E is a selected effect of X

The question from here becomes: Is the computer analogy (upon clarification) a good one? Are the functions of mind and brain analogously separable (such that the Autonomous Model might obtain in this case)?

7. Does functional separability obtain in the case of mind-brain?

For the Autonomous Model accurately to reflect what in fact occurs in some cases of mental disorder, the selected effects of the mind would have to (sometimes, at least) fail to be selected effects of the brain.

This simply cannot obtain. There is no way in which a mental process can be a naturally selected effect of the mind but not of the brain.[10] The only way in which a natural mental function (i.e., a genetically selected mental effect) can be configured into the mind through phylogenetic evolution is by being causally efficacious in the natural selection of the implementing organ—that is, the brain. To imply otherwise would be to succumb to dualism. It is not the case that the mind reproduces when the brain does not, nor that the brain dies while the mind lives on. Natural mental functions come about by brains yielding mental effects that confer a fitness advantage upon their organisms, thus causing the corresponding genotype to spread through the population. There is no alternative mechanism of action by which natural functions could arise. There is, in other words, no "mindware"-designer through which

[10] Or, strictly speaking, some other part of the body—but we have already conceded that all mental processes take place in neural tissues, and as such we shall not seriously consider this possibility.

https://doi.org/10.1017/psa.2022.7 Published online by Cambridge University Press

the mental derives distinct norms of operation. Mind-brain owes its structure and functional setup to a single process of evolution by natural selection acting upon the properties and characteristics of a single physical trait. In terms of intergenerational genetic selection, mind-brain develops as one.

Let us consider a concrete example. Fear is a mental effect. This is just to say that it is an effect, and that it is mental and phenomenal in nature. Fear plays a vital role in signaling danger, as well as in aiding our escape from predators and other threats through the activation of the sympathetic nervous system and is, as such, very plausibly naturally selected. In other words, fear is a natural mental function.

Uncontroversially, the structure and existence of the amygdalae are at least partly explained, in evolutionary terms, by their role in generating fear. Fear is, as such, a naturally selected effect of the brain. Moreover, there is no sense in which the "mental component" of fear was naturally selected independently of its physical implementation in the brain. The mind (in terms of intergenerational genetic transmission) lives and dies with the brain. We are accordingly not in a position to postulate some separate process of natural mental design, and thus separable mental functions.

This point bears emphasizing. The mere fact that "fear" is a mental or psychological effect does not imply that the brain was *not* naturally selected for the performance of it. Of course, the brain was designed not just for paradigmatic brain-things, like synaptic pruning or motor control, but for paradigmatic mental things, like mind-reading, emotion, or memory. The phenomenal character of mental processes obviously does not exclude them from being causally relevant to the natural selection of the organ that implements them. Indeed, there is no such relation.

In other words, failures of natural mental functions necessarily satisfy conditions 1 and 2 in relation to the brain. Natural mental dysfunctions are as such necessarily natural brain dysfunctions.

1) B instantiates ¬ME

And:

2) ME is a selected effect of B

Assuming that natural brain dysfunction is sufficient for brain disorder (as per the natural dysfunction analysis), it follows directly that legitimate mental disorders are necessarily brain disorders.

We can now reject Papineau's premise 3. Mind-brain are not analogous to software-hardware, and the AAMD fails to support its purported conclusion. Autonomous mental disorders (in the sense implied by the computer analogy) are a biological impossibility.

8. Objections

In this final section, I respond to three possible objections. First, in section 8.1, I consider whether ontogenetically selected mental effects can provide proponents

https://doi.org/10.1017/psa.2022.7 Published online by Cambridge University Press

of the AAMD with the kind of "functional separability" required to conform to the Autonomous Model. I briefly comment on why a proponent of the AAMD might not *want* to go down this road, before explaining why ontogenetically selected effects are, in any case, no threat to the argument I have advanced in this article. In section 8.2, I respond to the claim that multiply realized mental dysfunctions are ipso facto not brain dysfunctions (and thus not brain disorders). I argue that multiple realizability is unrelated to ascriptions of function and dysfunction in the brain, and indeed elsewhere. Finally, in section 8.3, I consider the implications of the argument I have provided for those theorists within the philosophy of medicine who reject the natural dysfunction analysis of mental disorder.

8.1 What about ontogenetically selected effects?

Ontogenetically selected mental effects are effects that are selected intraindividually through development, rather than intergenerational genetic selection. For example, my ability to recognize the letter B is a mental function but was selected through learning and lifetime ontogeny as opposed to genetic evolution. Could appealing to ontogenically selected effects give proponents of the AAMD the separability of mental and neural functions required by the Autonomous Model?

My response to this objection is two-pronged. I shall first argue that including ontogenetically selected effects within the category of natural functions threatens to undermine a central premise of the AAMD: the natural dysfunction analysis. I shall then argue that appealing to ontogenetically selected effects would not in any case salvage the AAMD, as ontogenetically selected effects are still selected effects of the brain and thus count as brain functions.

Firstly, it is not clear that failures of ontogenetically selected effects *should* be included in an analysis of medical disorder. Indeed, as I have defined disorder from section 2 onward (using Neander's version of the etiological theory, which appeals directly to selection at the level of the genotype), ontogenetically selected effects are excluded. They have also tended to be excluded from dysfunction-based accounts in philosophy of medicine traditionally (Boorse 1977, 2014; Wakefield 1992, 1999, 2014b; Matthewson and Griffiths 2018). Including failures of ontogenetically selected effects in the natural dysfunction analysis of disorder would thus require an additional theoretical move—a broadening of the notion of natural function to encompass lifetime-selected effects as well as genetically selected effects (see, e.g., Garson 2017, 2019).[11]

This move may not be favorable to the natural dysfunction analysis (and thus to the AAMD) for two sets of reasons: the possibility of (1) "healthy" ontogenetic *dys*functions and (2) "disordered" ontogenetic functions. As an example of the former, consider a woman who unlearns a lifetime of deeply ingrained bad habits. She is now failing to do something that was selected for through lifetime learning and ontogeny —but is she disordered? As examples of the latter: It seems possible that some

[11] I take the view, contra Garson's Generalised Selected Effects Theory (2017, 2019), that ontogenetically selected effects and genetically selected effects should be separated in a theory of function. I believe they play distinct roles and that forcing them to operate as one leads to contradictions and conflicts in functional norms. Expanding upon why would lead us too far afield, but this is another reason to be suspicious of any ontogenetic broadening of the natural dysfunction analysis.

https://doi.org/10.1017/psa.2022.7 Published online by Cambridge University Press

psychiatric disorders, such as phobias or OCD, may be selected through normal learning and conditioning. Recall the famous case of Watson and Rayner's Little Albert, who was conditioned to fear rats and other furry things (1920). Arguably an ontogenetically selected effect, but plausibly a disorder nonetheless.

Secondly, even were we to adopt a broadened version of the natural dysfunction analysis, ontogenetically selected mental effects still would not deliver the autonomous mental functions required by the Autonomous Model. The selection of mental processes through lifetime ontogeny is no less neural than their natural selection through evolution; in each case it is a matter a neural item (region, circuit, synapse, etc.) being selected *because* it yields some mental effect. Genetically selected mental effects, like the fight-or-flight response, are selected through the transmission of genes through differential intergenerational reproduction. Ontogenetically selected effects, like your ability to read English, are selected using mechanisms of neuroplastic adaptation (synapse selection, construction, changes to existing synaptic connections, repeated activation, or neurogenesis, to name some possible candidates) (Lillard and Erisir 2011; Garson 2019). If a neural item persists because it yields a mental effect, then that mental effect is a *selected effect of the brain*.[12] Selected effects of the brain are brain functions. Whether the implementing trait was favored by genetic or neural selection is of no consequence.

We can put this in terms of the conditions outlined in the preceding text. If (1) the failure of mental process ME (¬ME) is instantiated by the brain (B) and (2) ME is an ontogenetically selected effect of the brain, then ¬ME is an ontogenetic brain dysfunction. If ¬ME is an ontogenetic brain dysfunction, and ontogenetic dysfunctions count as medical disorders (per the broadened version of the natural dysfunction analysis, against which I have just cautioned), then ¬ME counts as a brain disorder. If ¬ME counts as a brain disorder, then ¬ME is *not* a candidate for autonomous mental disorder. Thus, the objection from ontogenetically selected effects fails.

But why aren't autonomous software functions, in the very same respect, functions of the realizing hardware?[13] Let us make the disanalogy between autonomous software functions and ontogenetic mental functions explicit. In modern general-purpose processers, software is a set of instructions written in programming language that tells any compatible hardware how to execute the software according to the specifications of its designer. When the software is installed onto the realizing hardware, the hardware changes its physical state in a number of ways—for example, electronic circuits are turned on and off. But this is no part of the software design as such. The autonomous software functions are fully fledged etiological artifact functions, quite independently of their eventual realization in the hardware, because they were selected for at the software writing stage—a stage that is both separate from and precedent to any physical state-changes in the hardware.

[12] See Garson (2019) and Garson and Papineau (2019) for a similar argument in defense of teleosemantics.

[13] I think there is a good sense in which autonomous software functions (contrary to ontogenetic and natural mental functions) really are *not* hardware functions (for reasons that I expand upon in this section). Note, however, that protestations to the contrary would not serve the proponent of the AAMD. If all software functions are ipso facto hardware functions, then the analogy from software malfunction to mental disorder does not even get off the ground; if there is no autonomous software dysfunction to begin with, then the issue of autonomous *mental* dysfunction does not even arise.

https://doi.org/10.1017/psa.2022.7 Published online by Cambridge University Press

In fact, the process of designing the software functions is often entirely abstracted away from the physical details of the hardware that later comes to implement them. When software is written in a so-called general purpose programming language (such as Python or Java) it can be implemented using a broad range of physical hardware, with the help of "compilers" that translate the general programming language into domain-specific instructions that can be understood by the specific hardware's central processing unit. As such, the software designer need not have had *any* particular state of hardware in mind when configuring effects such as "double spacing" (and similar software-specific functions) into the code at the software writing stage.

Contrast this with how ontogenetic mental functions come about. Ontogenetic mental functions are direct products of neuroplastic selection mechanisms acting upon the physical structure of the brain. That's it. There is no preneural writing of abstract instructions that could, even in principle, confer autonomous ontogenetic functions upon the supervenient mind.[14] This is the disanalogy. Autonomous software functions owe their status as functions, *not* to adaptations in the hardware, but to modifications to the software code prior to (and abstracted from) the software's physical realization. In mind-brain, however, there *is* no "mindware" (nor any precedent process of design)—just the developing brain, and a variety of neuroplastic changes throughout its ontogeny. There is simply no mind-brain analogue to the writing of software that could provide the etiological separability that obtains between software and hardware.[15]

In short, there is no more etiological separability in ontogeny then there is in evolution. Mind and brain, whether by evolution or by development, come about as one. It is simply not the case that ontogenetic selection acts on the "mindware" in isolation from the physical brain. As such, even if we resolve to sideline our earlier conceptual concerns, an appeal to ontogenetic selection is not going to deliver the kind of functional separability required by the Autonomous Model. If some particular ontogenetic mental function (such as the ability to speak Swahili or navigate a cab around London) has been configured into your mind, this is because the brain has adapted, through neuroplastic action, to perform that particular ontogenetic function. Accordingly, should this mental effect *fail*, that failure would still properly constitute a brain dysfunction.[16]

[14] Perhaps you are wondering, upon reading this, whether appealing to some other form of normativity, such as evolutive norms or cultural norms, might yield the separability required. I address this possible objection section 8.3.

[15] Do the switches and circuits of the hardware *inherit* the autonomous software functions upon having the software installed unto it? I don't think this is the case—that is, I don't think the intended functions of the instructions are imbued upon that which executes them. Consider an analogy. Imagine that Jane, the composer, designs a piece of piano music that is intended to serve as background music for a five-minute-long scene in a new film. Joe, the piano player, has received the sheet music and shows up for the first day of recording and performs the piece. The piece, it turns out, lasts only four minutes; Jane has made a mistake. Does Joe, by playing the piece, inherit the intended functions of the sheet music such that "being five minutes long" is not just a selected effect of the music piece but also of Joe, the piano player? Was Joe supposed to play for five minutes, even though the sheet music instructed him to play for four? It is hard to see how this could be the case. The intended duration of the piece is a selected effect of the sheet music, not of Joe. Joe's job is just to play the sheet music correctly.

[16] As a final nail in the coffin, consider that, even barring all the preceding, the proponent of AAMD (in pursuing this line of retort) would have to commit herself to the view that mental disorders are

https://doi.org/10.1017/psa.2022.7 Published online by Cambridge University Press

8.2 What about multiply realized mental dysfunctions?

Some in the literature mistake the possibility of multiple realized mental functions for evidence for autonomous mental disorder. The intuition is perhaps something like this: Even if the human brain was naturally selected to produce the mental function "fear," it still possible that different people's brains "do fear" in relevantly distinct ways. Surely, being multiply realized at the level of neurobiology, these kinds of effects do not properly constitute brain functions, nor the failure to produce them brain *dys*functions?

While I am happy to grant the in-principle possibility of multiply realized mental functions, as I shall go on to show, whether or not some mental dysfunction is multiply realized at the level of neural implementation is simply unrelated to whether it is, or is not, a brain dysfunction. My primary aim in this article has been to show that the computer analogy does not apply to the case of mind-brain, and thus that mental dysfunction without brain dysfunction (in the sense implied by the computer analogy) cannot obtain. A secondary aim has been to clarify how the argument for autonomous mental disorder works. Part of this entails theoretically distinguishing the AAMD from the supposition that mental dysfunctions may be multiply realized.

Jefferson has recently offered an argument that mental disorders should be understood as brain disorders if and only if they can be shown to track underlying neural regularities (2020). If a mental disorder is multiply realized at the level of neurobiology, then it is autonomous. Jefferson invokes the computer analogy but is of the mistaken view (along with others in this area, Papineau included) that the computer analogy illustrates an implication of nonreductive physicalism (see also Boorse 1976).

Indeed, as I hope to have established, the possibility of mental dysfunction in the absence of brain dysfunction depends on the applicability of the computer analogy to the case of mind-brain (which in turn depends on functional separability). If the analogy is truly apt, then mental dysfunction can occur in the absence of brain dysfunction, and mental disorder without brain disorder is possible. The possibility of multiply realized mental dysfunctions does not bear upon this at all. To see why, let us consider some less controversial cases of multiple realization.

The functional kind "corkscrew" can be realized in many distinct physical types (winged, lever, mounted, air-pressure, etc.). Whether failure to uncork bottles constitutes a dysfunction of some particular, say, winged corkscrew depends on whether that particular corkscrew (1) has uncorking bottles as a function and (2) is as a matter of fact failing to uncork bottles. Whether other artifacts also have uncorking as a function is of no relevance in determining this.

Consider convergent evolution. The functional kind "flight" is multiply realized in several distinct biological types (insects, birds, bats, etc). These distinct physical realizations were *all* naturally selected for flight and, as such, have "flight" among their

(generally speaking) characterized by failures of ontogenetically selected effects and brain disorders by failures of evolutionarily selected effects. This analysis fits poorly with usage, however. Forgetting one's native language, for example, is the failure of an ontogenically selected effect but seemingly more neurological than psychiatric. Similarly, it is not clear what ontogenetic function (if any) is violated in paradigm cases of mental disorder such as schizophrenia, bipolar and generalized anxiety. Perception, emotion and fight-or-flight are mental functions alright, but very plausibly naturally selected.

https://doi.org/10.1017/psa.2022.7 Published online by Cambridge University Press

natural functions. As such, should any one of these aerial organisms, such as a particular bat, be unable to perform flight, that would constitute a dysfunction of (some part of) that bat. The fact that other organisms have flight as a function, and as such may also be capable of instantiating failure to perform flight, is of no relevance whatsoever to whether some particular token bat is functioning as it should.

The implication for the brain is straightforward. Suppose that in Tom's brain "fear" is realized by some particular neural circuit N1. Suppose further that in Greg's brain "fear" is realized by some other neural circuit N2. Nevertheless, assuming N1 and N2 were selected for their performance of the mental function "fear," neural circuit N1 and neural circuit N2 still have "fear" as a function. As such, if N2 in Tom's brain were to fail to perform "fear," that failure would constitute a failure of neural circuit N2 to perform its function—that is, a dysfunction of N2.

Just as we agreed in the case of the corkscrew and the flightless bat, how "fear" is implemented in Greg's brain is entirely tangential to whether N2 in Tom's brain is functioning as it should. Whether failure to perform fear constitutes a dysfunction of N2 depends on (1) whether N2 is failing to perform "fear" and (2) whether N2 has "fear" as a function. Multiple realizability does not enter into the equation. Assuming, as I do, and as Jefferson indeed also does, that brain dysfunction is (per the natural dysfunction analysis) sufficient for brain disorder, it follows that mental disorders are brain disorders necessarily (whether or not they are multiple realized). Multiple realizability is tangential to the applicability of the computer analogy to the case of mental disorder, and it was simply a misunderstanding to think that the latter hinged on the former.

8.3 What of those who subscribe to a different theory of mental disorder?

The version of the AAMD that I have targeted thus far in this article proceeds explicitly from the natural dysfunction analysis of disorder, or something akin to it. There are, however, many theorists in the philosophy of medicine and psychiatry who invoke the analogy to software-hardware to disprove that mental disorders *are* brain disorders, but without anything like the natural dysfunction analysis in mind. Possible candidates here include Cooper (2007), Graham (2013), and Arpaly (2005).[17] I shall, in what follows, briefly sketch what I take to be the implications of the argument I have provided for those who reject a natural dysfunction analysis of mental disorder.

As noted, part of my objective in this article, beyond refuting the AAMD, has been to clarify what it is about the nature and functional setup of software and hardware that allows for the possibility of autonomous software dysfunction and, in turn, why it is so suggestive when invoked in psychiatric contexts. I have argued that autonomous software dysfunction depends on functional separability; not all software functions are hardware functions, not all software norms are hardware norms.

This analysis, I believe, still stands whether or not one accepts a natural dysfunction analysis of mental disorder. Anyone who claims that mental disorders are autonomous from brain disorders, *just like* software problems are autonomous from hardware problems in a computer, needs to be able to demonstrate that functional separability obtains in the case of mind-brain—whether one subscribes to a

[17] Thanks to two anonymous reviewers for pushing me on the issues addressed in this section.

https://doi.org/10.1017/psa.2022.7 Published online by Cambridge University Press

naturalist, a normativist or some other analysis of mental disorder. If functional separability does *not* obtain then mental disorders are *not* autonomous from brain disorders in the *same* way that software problems are autonomous from hardware problems in a computer. If so, the computer analogy is a red herring, and best left untouched.[18]

Having established this, the question thus becomes; would adopting a *different* analysis of mental disorder, in place of the natural dysfunction analysis, lead the proponent of AAMD down a more favorable path? Would it allow her to show that functional separability obtains, and that the computer analogy *does* apply? There are two main routes the proponent of AAMD, and enemy of the natural dysfunction analysis of mental disorder, might take. I shall examine the prospects for each in turn.

(1) Firstly, she could reject the natural dysfunction analysis wholesale. That is, she could adopt a single monist theory of medical disorder in place of the natural dysfunction analysis (e.g., one that is value-relative) and apply it to the body at large, including the mind and brain. For example, Cooper holds that a disorder is a "a bad thing to have, that is such that we consider the afflicted person to have been unlucky, and that can potentially be medically treated" (2002, 271). Cooper takes this analysis to apply across the board to somatic as well as mental conditions. What would this imply for the possibility of autonomous mental disorder?

Well, a mental disorder would be a bad, unlucky, mental thing that could potentially be medically treated. An *autonomous* mental disorder would be a bad, unlucky, and potentially treatable mental process realized by some neurobiological process that does *not* satisfy these criteria. But surely, if having some particular mental state is bad, unlucky, and in-principle subject to medical intervention, then this applies also to its underlying neural state? How could one be *unlucky* to have the mental state, but not the neural state that, as a matter of necessity, realizes it?[19]

Any proponent of the AAMD who accepts a *single* theory of mental and somatic disorder would have to answer a similar set of questions. As such, a nonnaturalist theory of disorder does not *in and of itself* do much to substantiate separability between mental norms of functioning and neural functions. However, there is another possible route that the proponent of AAMD could take: She could reject the natural dysfunction analysis only for *mental* disorder.

(2) The proponent of the AAMD could take the view that mental disorders and brain disorders are subject to distinct definitional criteria. For example, she could hold that brain disorders are constituted by natural dysfunction (the failure of the brain to yield one or more of its naturally selected effects, as already discussed) and that mental disorders, in contrast, are constituted by violations of evaluative, cultural, or personal norms. One possible line of reasoning here is to think of mental

[18] There may well be other ways of distinguishing paradigm mental disorders from paradigm brain disorders, but if there is no functional separability, mental disorders are *not* distinguished from brain disorders in that they are *like* software bugs in the "mindware" that the brain runs. For example, Bolton has suggested that paradigm brain disorders tend to be characterized by being comparatively less sensitive to forms of psychosocial intervention (Bolton 2013).

[19] Remember that, even per nonreductive physicalism, the underlying neural state is *sufficient* for producing the supervenient mental state. There is accordingly no scenario in which you could instantiate the underlying neural state without it realizing the disordered mental state. It is as such hard to see how the "bad luck" in having the mental state could come apart from its neural implementation.

https://doi.org/10.1017/psa.2022.7 Published online by Cambridge University Press

disorders as infringements on norms that have arisen through the intergenerational transmission of traits by cumulative cultural evolution (see Boyd and Richerson 1996, 2005).[20]

In contrast to (1), (2) would provide functional separability. Because the norms of functioning to which the mind is subject, on this view, are distinct and separable from the naturally selected brain functions, one could have mental dysfunction (say, failure of the mind to conform to some evaluative norm of what it is good to feel or think) without any brain dysfunction (in the absence of any infringement on the evolutionarily conditioned norms of functioning against which the brain is judged). I readily concede that the argument provided thus far in this article has no real teeth against this view. Following strategy (2), mental disorder without brain disorder *is* possible. Before closing however, I shall outline two reasons to doubt the viability—or at least, attractiveness—of this position.

Firstly, the proponent of position (2) would have to concede that she is really talking about two different things. The claim that "mental disorder does not entail brain disorder" ought really to be read as "mental disorder[1] does not entail brain disorder[2]." This is less surprising prima facie and less interesting in substance. Some of the attraction of the AAMD as advanced by Papineau and Wakefield was that it purported to show how the mind could become disordered on the very same terms as the brain, but without the brain states instantiating corresponding disorder. It is somewhat less intriguing to find out that all brain processes doing what they were designed for by evolution is compatible with a mental process simultaneously failing to conform to some entirely distinct norm. We should have expected this to be the case at the outset.

Secondly, in accepting a distinct definitional criterion for mental disorder, unrelated to that which defines somatic disorder, the proponent of the AAMD is now left open to the antipsychiatric challenge the argument was originally conceived to rebut: that mental disorders are not *real* disorders. Suppose that *real* disorder—in the mind of a critic of psychiatry—means disorders in the same sense as, or at least with some reasonably strict analogy to, paradigm neural and somatic disorders. The proponent of the AAMD would have to concede that mental disorders are *not real* in this sense—precisely because she employs distinct criteria in the mental and the somatic cases. On her view, mental disorders share nothing definitionally in common with somatic and biomedical disorders. They are not the same kind of thing.

There is nothing inconsistent about strategy (2), and it does provide for autonomous mental dysfunction. However, it is less interesting and less surprising than the AAMD as originally presented, and it fails to solve one of the key theoretical problems which the argument was originally introduced to remedy. In short, you may get autonomous mental disorders—but at what cost?

[20] I have said that ontogenetically selected effects cannot serve to substantiate the AAMD, and now I am about to say that culturally selected effects perhaps could, so (to avoid confusion) let me be clear about precisely what I take the difference to be. I take ontogenetically selected effects to be effects that have caused their corresponding traits to have been selected on a developmental timescale intraindividually. Neural selection is the paradigm example, but the selection of antibodies is another possible case (Garson 2019). Culturally selected effects, on the contrary, are selected, usually intergenerationally, but at least interindividually within groups that can be said to make up "cultures," for example, the intergenerational transmission of gendered practices (see Godman 2018).

https://doi.org/10.1017/psa.2022.7 Published online by Cambridge University Press

9. Concluding remarks

I have argued that the AAMD fails to support its stated conclusion. Because the selection histories that give rise to the functions of software-hardware are distinct and separable, the functions of software-hardware are also separable. This is what provides for the possibility of software problems in the absence of hardware problems or "autonomous software dysfunction." Mind-brain, however, develops as one, subject to the same selection pressures. There is as such no analogous disassociation of mental functions from brain functions, and the AAMD thus fails.

The possibility of autonomous software dysfunction does not extend to imply the possibility of autonomous mental disorder as has traditionally been purported in philosophy of medicine and psychiatry. In fact, the natural dysfunction analysis of disorder, as defended by Wakefield, Papineau, Neander, and others, implies a view of mental disorders as brain disorders necessarily; not because mental dysfunctions are *reducible* to types of neurophysiological abnormality, but because natural mental functions are naturally selected effects of the brain, in and of themselves. Thus, should a natural mental function fail, that failure would be a brain dysfunction.

Acknowledgments. I owe a special debt of gratitude to David Papineau for his support at every stage of producing this research. I am also deeply grateful to Alexander Bird and Michael Pauen for invaluable feedback and support especially on earlier drafts. I would also like to warmly thank Anneli Jefferson, Cecily Whiteley, participants at the work-in-progress seminars at King's College London and Berlin School of Mind and Brain, and two anonymous reviewers to this journal for insightful comments and questions. I am indebted as well to Simon Lord and Geronimo Hampton for permitting me to share their office. This research was supported by the London Arts and Humanities Partnership and the Deutscher Akademischer Austauschdienst.

References

Arpaly, Nomy. 2005. "How It Is Not 'Just Like Diabetes': Mental Disorders and the Moral Psychologist." *Philosophical Issues* 15 (1):282–98. https://onlinelibrary.wiley.com.

Bolton, Derek. 2013. "Should Mental Disorders Be Regarded as Brain Disorders? 21st Century Mental Health Sciences and Implications for Research and Training." *World Psychiatry* 12 (1):24–25.

Boorse, Christopher. 1976. "What a theory of mental health should be." *Journal for the Theory of Social Behaviour* 6 (1):61–84.

Boorse, Christopher. 1997. "Health as a Theoretical Concept." *Philosophy of Science* 44 (4):542–73.

Boorse, Christopher. 2014. "A Second Rebuttal on Health." *Journal of Medicine and Philosophy* 39 (6): 683–724.

Borsboom, Denny, Angélique O. J. Cramer, and Annemarie Kalis. 2019. "Brain Disorders? Not Really: Why Network Structures Block Reductionism in Psychopathology Research." *Behavioral and Brain Sciences* 42:1–12. https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/brain-disorders-not-really-why-network-structures-block-reductionism-in-psychopathology-research/D5A20455723B237C60E379D29F8797B1.

Boyd, Robert, and Peter J. Richerson. 1996. "Why Culture Is Common, but Cultural Evolution Is Rare." *Proceedings-British Academy* 88:77–93. https://www.thebritishacademy.ac.uk/documents/3949/88p077.pdf.

Boyd, Robert, and Peter J. Richerson. 2005. *The Origin and Evolution of Cultures*. Oxford: Oxford University Press.

Brülde, Bengt, and Filip Radovic. 2006. "What Is Mental about Mental Disorder?" *Philosophy, Psychiatry, & Psychology* 13 (2):99–116.

Coelho Mollo, Dimitri. 2021. "Against Computational Perspectivalism". *The British Journal for the Philosophy of Science* 72 (4):1129–1153. https://www.journals.uchicago.edu/doi/full/10.1093/bjps/axz036.

https://doi.org/10.1017/psa.2022.7 Published online by Cambridge University Press

Cooper, Rachel. 2002. "Disease." *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 33 (2):263–82.
Cooper, Rachel. 2007. *Psychiatry and Philosophy of Science*. Stocksfield: Acumen.
Cooper, Rachel. 2017. "Health and Disease." In *The Bloomsbury Companion to Contemporary Philosophy of Medicine*, edited by James A. Marcum, 275–96. London: Bloomsbury Academic.
Dewhurst, Joe. 2014. "Mechanistic Miscomputation: A Reply to Fresco and Primiero." *Philosophy & Technology* 27 (3):495–98.
Fresco, Nir, and Giuseppe Primiero. 2013. "Miscomputation." *Philosophy & Technology* 26 (3):253–72.
Garson, Justin. 2017. "A Generalized Selected Effects Theory of Function." *Philosophy of Science* 84 (3):523–43.
Garson, Justin. 2019. *What Biological Functions Are and Why They Matter*. Cambridge: Cambridge University Press.
Garson, Justin, and David Papineau. 2019. "Teleosemantics, Selection and Novel Contents." *Biology & Philosophy* 34 (3):1–20.
Godman, Marion. 2018. "Gender as a Historical Kind: A Tale of Two Genders?" *Biology & Philosophy* 33 (3): 1–16.
Graham, G. 2013. "Ordering Disorder: Mental Disorder, Brain Disorder, and Therapeutic Intervention." In *The Oxford Handbook of Philosophy and Psychiatry*, edited by K. W. M. Fulford, Martin Davies, Richard G. T. Gipps, George Graham, John Z. Sadler, Giovanni Stanghellini, and Tim Thornton, 512–30. Oxford: Oxford University Press
Insel, Thomas, Bruce Cuthbert, Marjorie Garvey, Robert Heinssen, Daniel S. Pine, Kevin Quinn, Charles Sanislow, and Philip Wang. 2010. "Research Domain Criteria (RDoC): Toward a New Classification Framework for Research on Mental Disorders." *American Journal of Psychiatry* 167 (7):748–51.
Insel, Thomas R., and Bruce N. Cuthbert. 2015. "Brain Disorders? Precisely." *Science* 348 (6234):499–500.
Jefferson, Anneli. 2020. "What Does It Take to Be a Brain Disorder?" *Synthese* 197 (1):249–62.
Kingma, Elselijn. 2013. "Naturalist Accounts of Mental Disorder." In *The Oxford Handbook of Philosophy and Psychiatry*, edited by K. W. M. Fulford, Martin Davies, Richard G. T.
Lillard, Angeline S., and Alev Erisir. 2011. "Old Dogs Learning New Tricks: Neuroplasticity beyond the Juvenile Period." *Developmental Review* 31 (4):207–39.
Matthewson, John, and Paul E. Griffiths. 2018. "Evolution, Dysfunction, and Disease: A Reappraisal." *The British Journal for the Philosophy of Science* 69 (2):301–27.
Millikan, Ruth Garrett. 1989. "In Defense of Proper Functions." *Philosophy of Science* 56 (2):288–302.
Neander, Karen. 1983. "Abnormal Psychobiology." PhD diss., La Trobe University.
Neander, Karen. 1991. "Functions as Selected Effects: The Conceptual Analyst's Defense." *Philosophy of Science* 58 (2):168–84.
Papineau, David. 1994. "Mental Disorder, Illness and Biological Disfunction." *Royal Institute of Philosophy Supplements* 37:73–82. https://www.cambridge.org/core/journals/royal-institute-of-philosophy-supplements/article/mental-disorder-illness-and-biological-disfunction/A402FDA35E2649F816A3A5325B1668C9.
Svensson, Tommy. 1990. "On the Notion of Mental Illness: Problematizing the Medical-Model Conception of Certain Abnormal Behaviour and Mental Afflictions." PhD diss., Linköpings Universitet.
Szasz, Thomas S. 1960. "The Myth of Mental Illness." *American Psychologist* 15 (2):113–18.
Tucker, C. 2018. How to Explain Miscomputation. *Philosopher's Imprint* 18 (24):1–17.
Wakefield, Jerome C. 1992. "The Concept of Mental Disorder: On the Boundary between Biological Facts and Social Values." *American Psychologist* 47 (3):373–88.
Wakefield, Jerome C. 1999. "Evolutionary Versus Prototype Analyses of the Concept of Disorder." *Journal of Abnormal Psychology* 108 (3):374–99.
Wakefield, Jerome C. 2006. "What Makes a Mental Disorder Mental?" *Philosophy, Psychiatry, & Psychology* 13 (2):123–31.
Wakefield, Jerome C. 2014a. "Wittgenstein's Nightmare: Why the RDoC Grid Needs a Conceptual Dimension." *World Psychiatry* 13 (1):38–40.

https://doi.org/10.1017/psa.2022.7 Published online by Cambridge University Press

Wakefield, Jerome C. 2014b. "The Biostatistical Theory Versus the Harmful Dysfunction Analysis, Part 1: Is Part-Dysfunction a Sufficient Condition for Medical Disorder?" *The Journal of Medicine and Philosophy: A Forum for Bioethics and Philosophy of Medicine* 39 (6):648–82.

Wakefield, Jerome C. 2017. "Addiction and the Concept of Disorder, Part 2: Is Every Mental Disorder a Brain Disorder?" *Neuroethics* 10 (1):55–67.

Wakefield, Jerome C. 2020. "Addiction from the Harmful Dysfunction Perspective: How There Can Be a Mental Disorder in a Normal Brain." *Behavioural Brain Research* 389:1–9. https://doi.org/10.1016/j.bbr.2020.112665.

Watson, John B., and Rosalie Rayner. 1920. "Conditioned Emotional Reactions." *Journal of Experimental Psychology* 3 (1):1–14.

Cite this article: Fagerberg, Harriet. 2022. "Why Mental Disorders are not Like Software Bugs." *Philosophy of Science* 89 (4):661–682. https://doi.org/10.1017/psa.2022.7