Acta Analytica
https://doi.org/10.1007/s12136-024-00587-2

Identity Theory and Falsifiability

Anders Søgaard[1]

Received: 17 August 2023 / Accepted: 5 February 2024

Abstract
I identify a class of arguments against multiple realization (MR): *Bookof Sand* arguments. The arguments are in their general form successful under reasonably uncontroversial assumptions, but this, on the other hand, turns the table on identity theory: If arguments from MR can always be refuted by *Bookof Sand* arguments, is identity theory falsifiable? In the absence of operational demarcation criteria, it is not. I suggest a parameterized formal demarcation principle for brain state/process types and show how it can be used to identify previously unconsidered contenders for evidence for MR, e.g., binary classification, division, and sorting. For these to be *actual* instances of MR, the corresponding psychological kinds must be verifiably, relevantly similar. I also briefly discuss possible linguistic, behavioral, and experimental demarcation criteria for psychological kinds.

1 Introduction

Multiple realization (MR)[1] refers to the belief that the same psychological *kind* can be realized by significantly different types of brain states or processes. Pain or anger or recognizing a cat can, in other words, be implemented in significantly different ways. Putnam (1960) presented MR as a counter-argument to so-called identity theory. He took identity theory to be the hypothesis that for each psychological kind, there is a unique physical kind that is "nomologically coextensive" with it. That is, by necessity, the occurrence of one entails the other. Many empirical phenomena have been presented as supposed evidence for MR. Two commonly cited example classes

[1] Polger and Shapiro (2016) present two definitions of multiple realization: (a) Multiple realization occurs if and only if two (or more) systems perform the same function in different ways, and (b) multiple realization occurs if and only if two (or more) systems perform relevantly the same function in relevantly different ways. Some observations will refute only the basic definition; other observations will refute both.

✉ Anders Søgaard
soegaard@di.ku.dk

[1] Department of Computer Science, Department of Communication, and Pioneer Center for Artificial Intelligence, University of Copenhagen, Lyngbyvej 2, DK-2100 Copenhagen, Denmark

Published online: 26 February 2024

are (i) similar psychological kinds across species with very different neural layout, e.g., pain in humans and fish, and (ii) neural plasticity (Figdor, 2010; Polger & Shapiro, 2016; Michel, 2019).

The crux of the MR debate is whether *relevantly similar* psychological kinds can be realized by *relevantly different* brain state/process types. I will focus on one class of arguments against MR. The general argument or argument scheme, which I will call *Book of Sand* arguments—and motivate at more length in Section 2—goes as follows:

> **Book of Sand** (ϕ). *While ϕ is presented as evidence for MR, i.e., that a psychological kind is realized by two relevantly different brain state/process types, ϕ fails to prove MR, either because the brain state/process types realize different psychological kinds, or because the brain state/process types are not relevantly different.*

Arguments of this form appear in Polger and Shapiro (2016) and Michel (2019). The arguments appeal to uncertainty about what *exactly* the necessary conditions are for something to be of a particular brain state/process type or of a particular psychological kind. What, exactly, makes two brain state types relevantly different? Or how about two psychological kinds—when exactly is it relevant to discriminate between the two? The arguments provide support for the view that MR is unjustified, and as such, the arguments can be seen as a defense of identity theory. The second part of my article, however, turns the table on identity theory.

For would it ever be possible to find empirical support for MR, with *Book of Sand* arguments around? If not, identity theory is, to the extent it reduces to the negation of MR, unfalsifiable. Identity theory says that each mind state type (psychological kind) μ_i is realized by a brain state type β_i. *Book of Sand* arguments trade on uncertainties around the necessary conditions for type membership. If membership of μ_i is underdetermined, or membership of β_i is underdetermined, it seems *impossible* to prove that relevantly similar psychological states (members of μ_i) can be realized by relevantly different brain state/process types, β_i and β_j with $\beta_i \neq \beta_j$? For how can one show that "hunger" is relevantly similar for mammals and vertebrates (being members of the same psychological kind), for example, but realized by relevantly different brain state/process types, if the necessary conditions for membership are not entirely known? Can we not always say that the appearance of MR is explained away by the two brain states or processes in fact being members of the same type, or by the two psychological kinds being different at some level, by some previously unidentified condition?

What can we do to settle the score? The falsifiability of identity theory depends on *both* our definition of the extension of brain state/process types *and* the ability of psychology to ground psychological kinds in observations, e.g., behavior or experimental data. The extension of brain state/process types is maybe easiest to settle, e.g., using biological or mathematical criteria, and I will propose a second-order similarity metric to this end, relying on ϵ-isometry. Grounding psychological kinds is trickier, and after

briefly reviewing classic options (linguistic, behavioral, and experimental demarcation criteria), I leave it as an open question in what sense this is at all possible.[2]

2 Book of Sand Arguments

Jorge Luis Borges's *Book of Sand*[3] is a (fictitious) book with an infinite number of pages. One type of argument against MR is to say there are more pages in the *Big Book of Science* than has previously been considered. Say the book describes two psychological kinds as similar, but on the previously unseen pages, properties of the two psychological kinds are described that differentiate the two. This refutes MR, because the different brain state/process types realize different psychological kinds, after all. *Book of Sand* arguments work in the opposite direction, too: Say the book describes two brain state/process samples in ways that would at first lead us to think they belonged to different kinds. On the unseen pages, now, we find evidence that what seems to be properties that differentiate the two, are in fact not discriminatory. The two samples belong to the same brain state/process type, after all.

The two classic arguments against MR are the Grain Argument (Bechtel, 1999) and the Causal Powers Argument (Kim, 1992). The Grain Argument says that MR is the result of analyzing psychological kinds and brain state/process types at different levels of granularity. Philosophers think of psychological kinds in a coarse-grained fashion, whereas more fine-grained criteria are used to identify brain state/process types. This seems somewhat related to *Book of Sand* arguments. The Causal Powers Argument, on the other hand, is orthogonal to *Book of Sand* arguments: It refers to idea that structure-independent psychological kinds are not causal kinds and therefore do not count as scientific kinds in the first place.

How do *Book of Sand* arguments relate to the Grain Argument? The Grain Argument has a much more limited scope. Both claim that MR dissolves when we compare psychological kinds and brain state/process types more carefully. The Grain Argument says readjustment amounts to moving from one level of analysis to another. *Book of Sand* arguments are, in a sense, more agnostic, saying ambiguity arises from disagreement about what properties are relevant, and how to measure similarity.

Book of Sand arguments have, for example, been applied to the case of fish pain: **Do Fish Feel Pain?** Michel (2019) considers the so-called *no cortex, no cry* argument, which goes as follows. If animals feel pain, they have a neocortex ($p \rightarrow q$). Fish do not have a neocortex ($\neg q$). Therefore (by *modus tollens*), fish do not feel pain ($\neg p$). The question, of course, is whether the first premise ($p \rightarrow q$) is true. Is a neocortex a necessary condition for feeling pain? The premise at least *seems* true for a human. If the premise is not true, on the other hand, and if fish do feel pain, this is evidence

[2] Philosophers of science—including readers of Quine, Kuhn, Lakatos, and others—will know that falsifiability is not an uncontroversial demarcation principle. Many unscientific claims are falsifiable, and many scientific claims are not directly falsifiable, e.g., Newton's first law. Our discussion of falsifiability should be uncontroversial, however, since what is at stake is whether a prominent debate, between proponents of identity theory and multiple realization, can ever, in principle, be decided. Running in circles is unequivocally suboptimal for a scientific community.

[3] The story (*El libro de arena*) is last of 13 stories by the Argentinian writer, in a book of the same name. The first English translation was published in The New Yorker.

for MR. If brain state/process types in a neocortex are relevantly different from brain state/process types elsewhere, that is. If we assume MR, the absence of a neocortex in fish is not a valid reason to reject fish pain, because the first premise of the *no cortex, no cry* argument is no longer assumed to be true. The MR response to this argument is a prominent one, as shown by Michel. Now what would *Bookof Sand* arguments against this form of MR look like? Well, we have two strategies: We can either hope to find sufficient similarities between the brain state/process types that realize pain in humans and pain in fish. That is, we establish biological or mathematical criteria such that the brain state/process types that realize human and fish pain fall into the same bucket. Or we can hope to find functional differences between the corresponding psychological kinds, i.e., ways in which pain is functionally different across the two species. This assumes that we have a functional definition of pain in the first place, of course.

Bookof Sand arguments are also found in recent discussion of the classic case of the rewired ferrets:

Rewired Ferrets Sharma et al. (2000) rewired thalamocortical connections in ferrets to make cortical areas serve new processing tasks. Shapiro (2004) and Polger (2009) disagree the ferrets count as evidence for MR: The ferrets only achieve a very limited kind of vision, and the newly rewired auditory cortex became organizationally more similar to the primary visual cortex. Both arguments are *Bookof Sand* arguments.

There's more out there: Premack (2007), for example, goes through eight examples of psychological kinds alleged to be shared between humans and other species, arguing that across the board, when you look into the specifics, dissimilarities accumulate faster than similarities. Or construct your own: Bats and whales both use echolocation, an instance of convergent evolution. So multiple realization? Not so fast. For consider now the differences between the two forms of echolocation. Bats, for example, create their sonar pulses using their voicebox, while whales pass air through their nasal bones. Oh, this piece of evidence goes to the identity theory pile? Maybe. On the other hand, bats and whales use the same strategies for detecting the rebounding echoes, and more or less the same genes seem to be involved.

Examples such as fish pain and rewired ferrets show how the scope of the Grain Argument is too limited. None of these arguments against MR are, strictly speaking, instances of the Grain Argument. Shapiro and colleagues all argue that the brain state/process types in question were *not* relevantly different, after all, *or* the psychological kinds in question were, but none of them argue that a move from one level of analysis to another, will give us a one-to-one mapping between psychological kinds and brain state/process types. *Bookof Sand* arguments are agnostic about levels of analysis, and for it to work, you do *not* need to be able to identify such levels, making it stronger than the Grain Argument.

Why are *Bookof Sand* arguments so pervasive? *Bookof Sand* arguments trade on the vagueness of kinds or types. Kinds or types are governed by stabilizing functions (Shea, 2018). Genotypes, for example, are reinforced by replication. Flamingos (*Phoenicopteridae*) form an easily identifiable natural kind, because flamingos inherit their properties through replication. Mutation leads to occasional drift, and sometimes such drift is sufficient to motivate *Bookof Sand* arguments, but most of the time

genotypes orbit around equilibria. Cultural kinds, e.g., a kind of dance, are governed by another stabilizing function, namely imitation learning (Shea, 2018). Innovation, again, leads to drift, but the pressure of imitation learning makes tango recognizably tango over extended periods of time.

No such stabilizing functions govern brain state types and psychological kinds. Brain states do not seem to replicate. And even if some psychological kinds are imitation learned—such as surprise, disgust, or anger (Ekman et al., 1983)—we typically do not have explicit satisfaction criteria for when psychological kinds have been learned. While it is easy to jot down a recipe for cooking a pizza or dancing a tango, it is much harder to define exactly what goes into feeling in love or the subjective experience of fear.

One possible objection is that brain states may converge toward specific configurations for reasons unknown to us. Brains may, for example, minimize free energy and learn optimal models of the environment in their attempt to control it. Or brains may fold like proteins to minimize certain loss functions and so on. Whether such principles of organization would lead to brain states orbiting around a predefined set of brain state types or configurations of basic types is, I contend, an empirical question. My business here is merely pointing out how *Bookof Sand* arguments trade on the fact that we do not know the answer to this empirical question—at least not in detail.

How about mirror neurons? Mirror neurons located in the premotor cortex and the inferior parietal cortex—first studied by Rizzolatti and Craighero (2004)—activate both when an individual performs an action and when they observe the same action being performed by someone else. In other words, these neurons "mirror" the behavior of others as if the observer themselves were performing the action. Is this not a form of replication of brain states? The significance of mirror neurons is still an area of ongoing research, and unless the replication is accurate enough to induce distributions around "genotypical" brain state types, it will not help resolve questions around MR.

3 Unfalsifiability Proof

Assume that brain state type β_i is described by a finite set of atomic propositions $\mathcal{B}_i$, and brain state type β_j is described by a finite set of atomic propositions $\mathcal{B}_j$. What we refer to as mind state type μ_i can be described by a finite set of atomic propositions $\mathcal{M}_i$, of which a subset $\mathcal{M}_i^c$ are constitutive. What we refer to as mind state type μ_j is described by a finite set of atomic propositions $\mathcal{M}_j$, of which a subset $\mathcal{M}_j^c$ are constitutive. Constitutive atomic propositions correspond to the necessary conditions for a state to belong to this state type; they are, so to speak, the *relevant* propositions.

You can think of β_i and β_j as realizations of hunger (or pain or the recognition of an oak tree) in two different species, for example. Now, μ_i is hunger in an elephant; μ_j is hunger in a cat. If $\mu_i = \mu_j$ (or $\mu_i \sim_r \mu_j$, where $\sim_r$ means relevantly similar), hunger is evidence for MR.

That is, two conditions must be satisfied for there to be evidence for MR:

(p_1) μ_i and μ_j must be relevantly similar.

(p_2) β_i and β_j must be relevantly different.

In *Book of Sand* arguments, either μ_i and μ_j are shown *not* to be relevantly similar, or β_i and β_j are shown *not* to be relevantly different. The arguments for saying μ_i and μ_j are different ($\neg p_1$), after all, tend to revolve around functional differences, but they can also be experiential, behavioral, or experimental. Consider, for example, the broad pallet of methods used in psycho-metrics. The arguments for saying β_i and β_2 are similar, after all, also vary: In the case of *situs inversus viscerum*, β_i and β_2 are mirror images. Fly and cow opsins are homologous. Both isomorphism and homology can, according to Polger and Shapiro, be used to establish $\neg p_2$. If all evidence for MR can be refuted one way or the other, identity theory, as a consequence, clears its case (for now) and, all things being equal, becomes preferable over MR by Occam's razor.

Our refutation was, of course, too easy. In the absence of more precise criteria for how we can establish $\neg p_1$ and $\neg p_2$, identity theory becomes *unfalsifiable* by MR.[4] To see this, let $\mathcal{M}_i^{\bar{c}} \subset \mathcal{P}$ and $\mathcal{M}_j^{\bar{c}} \subset \mathcal{P}$ be the complements of the sets of constitutive atomic propositions for our two mind states (hunger in elephants and hunger in cats). We can now make the following observations:

Mind State Type Principle If $\mathcal{M}_i^{\bar{c}} \backslash \mathcal{M}_j^{\bar{c}} \neq \emptyset$, i.e., there is at least one proposition that is non-constitutive of elephant hunger, that is not non-constitutive of cat hunger, and if we can freely add a proposition $p \in \mathcal{M}_i^{\bar{c}}$ to $\mathcal{M}_i^c$, we will always be able to ensure $\mathcal{M}_i^c \neq \mathcal{M}_j^c$.

To see why this is so, consider how $\mathcal{M}_i$ and $\mathcal{M}_j$ are non-identical. Elephants or cats differ in many ways, even when they both feel hunger. This is our first premise ("there is at least one proposition that is non-constitutive of elephant hunger, but constitutive of cat hunger"). If it is not clear what exactly is constitutive of hunger in elephants, it becomes trivial to make the two mind states relevantly different. In other words, if we can always find an atomic proposition p that is true, but non-constitutive for μ_i and false for μ_j, a constitutive atomic proposition for μ_i, we can always make the two mind state types relevantly different.

The converse principle holds for brain state types:

Brain State Type Principle If we can freely add an atomic proposition $p \in \mathcal{B}_i^c$ to $\mathcal{B}_i^{\bar{c}}$, or an atomic proposition $p \in \mathcal{B}_j^c$ to $\mathcal{B}_j^{\bar{c}}$, we will always be able to ensure $\mathcal{B}_i^c = \mathcal{B}_j^c$.

Note that any of the principles is sufficient to establish the unfalsiability of identity theory. In the next section, we will consider ways of mitigating the unfalsifiability of identity theory.

First, some preliminary housekeeping: Since scientific descriptions of brain state/process types and psychological kinds may be incomplete, the key for establishing MR is exhausting (emptying out) $\mathcal{M}_i^{\bar{c}} \cup \mathcal{M}_j^{\bar{c}}$ and $\mathcal{B}_i^c \cap \mathcal{B}_j^c$. The latter should in theory be easiest, for this is a small set of established sufficient conditions. The trouble, of course, is that when, say, homology is used as an argument for similarity in all relevant respects, this argument is not explicitly referencing specific now-to-be-seen-as-non-constitutive atomic propositions. Our first proposal for making identity theory falsifiable will therefore be to require that $\neg p_2$ can only be established through direct

[4] If identity theory is not falsifiable by MR, it does not follow that it is not falsifiable *simpliciter*. It could be falsified by something other than MR.

observations, referring to the specific properties in question. As for the former set of atomic properties, i.e., $\mathcal{M}_i^{\bar{c}} \cup \mathcal{M}_j^{\bar{c}}$, the trouble is that this set is too open-ended. We will therefore have to ensure the falsifiability of identity theory by restricting the set of relevant atomic propositions that can be used to differentiate two mind state types. Let $\mathcal{R} \subseteq \mathcal{P}$ be the set of all relevant propositions, i.e., the propositions that would constitute relevant differences between two mind state types. In the next section, we will discuss possible necessary conditions for *relevant* atomic propositions.

4 Making Identity Theory Falsifiable

Fish pain and rewired ferrets illustrate how identity theory and multiple realization are in bad need of effective demarcation criteria. Below, I suggest possible criteria. I exclude two criteria up front: qualia and homology. My reasons are as follows: The opposite of homologous organs is analogous organs which do similar jobs in two taxa, but which evolved separately and were not present in their most recent common ancestor. Homology is of no help to us, however, since the definition of homology depends on how functional kinds are defined, and when we define a physical realization to be present or not.[5] Homology, in other words, has as much explaining to do as identity theory itself. Qualia is subjective per definition and will not help us define operational demarcation criteria. Neither qualia nor homology comes with explanatory value for demarcation of kinds. Qualia and homology are not the only poor demarcation criteria in the MR debate, but they are arguably the most popular ones.

I will instead introduce a second-order distance metric to demarcate brain state/process types and argue for its possible adequacy,[6] I will then show how a derived demarcation principle can be used to identify novel, interesting contenders for MR:

For a moment, consider MR in neural networks rather than in mammals or vertebrates. What does it mean for a psychological kind to be multiply realized in neural networks? In a neural network, a brain state/process type is also a set of neural activations, not in biological tissue, but in digital code. Imagine you pass two neural networks monocolor images of all sorts and ask yourself whether the neural networks encode color in relevantly similar ways. For each network and each color, you extract a set of vectors of neural activations, e.g., a vectorization of all activations, or just the activations at the outer layer. You can now ask: Is the vector space induced by the first neural network for color ϵ-isometric (see below for definition) to the vector space

[5] To see this, consider a homeostatic property cluster account (Boyd, 1989) of natural kinds such as birds and bats, for example. Let $c^i = \{p_1^i, \ldots, p_n^i\}$ be the properties commonly associated with birds, and $c^a = \{p_1^a, \ldots, p_n^a\}$ the properties commonly associated with bats. Empirically, bird wings are both homologous and *not* homologous to bat wings. Bird wings are homologous to bat wings as derivatives of forelimbs, but, on the other hand, bird wings are not homologous to bat wings as wings, because the forelimbs of the common ancestors of birds and bats were not (what we would call) wings. So, we have traded the question of whether wings are multiply realized in birds and bats, for the question of whether wings are multiply realized in birds or bats and their common ancestors.

[6] Many other metrics could probably do the job just as well. Representational similarity analysis (Kriegeskorte et al., 2008), for example, is an established metric for comparing brain imaging data, with very similar properties.

induced by the second network? This strategy is used in Li et al. (2023) to evaluate whether language models and humans encode relevantly similar world knowledge, and before that, in studies of cross-lingual and multi-modal similarities between language models.

In mathematics, a quasi-isometry is a function between two metric spaces that respects the large-scale geometry of these spaces and ignores their small-scale details. The concept was introduced by mathematician Mikhael Gromov and builds on the concept of isometry, which refers to distance-preserving metric space transformations. Let us say we represent brain state/process types as a set of vectors in space (or equivalently, a densely connected, weighted graph). We can think of brain image vectors as approximations or representations of the underlying neural activation vectors.[7] Brain state/process types are thus, on this view, sets of neural activation (brain image) vectors. Two brain state/process types s and t are isometric if and only if there is a map from s to t, such that the vectors of s map to the vectors of t, and the distance between the vectors of s is up to the additive constant ϵ within a factor of the distance between the corresponding vectors of t. ϵ thus parameterizes this demarcation principle for distinguishing relevantly different brain state/process types.

Our new demarcation principle for brain state/process types can be used to identify relevantly different realizations of what seems to be similar psychological kinds, but also to rule out cases in which realizations only seem different, such as in the color-encoding neural networks above. It should be easy enough to see how isometry can also give us demarcation criteria for human brain state/process types. For each brain state/process type, you collect or approximate the set of neural activation vectors, say by averaging the brain imaging vectors of multiple subjects or trials. You now compute the cosine distances between them and ask if the resulting spaces are ϵ-isometric. Note, by the way, how such a metric only makes sense on a Hebbian worldview. If a proponent of Language of Thought thinks the state of feeling pain amounts to having $p =$"I have pain" in your belief box or in your knowledge box, MR is more easily defined in terms of logical subsumption.

The problem with mind states such as hunger, pain, or recognizing an oak tree is that we currently have no way to tell exactly which propositions are constitutive of these mind states. Binary classification, division, and sorting, however, may be better examples, because we know relevantly different algorithms for solving these tasks and can quantify their differences, irrespective of how they are implemented in different substrates.

I will discuss three examples below. They are all examples of problems for which different algorithms implement equivalent functions, and presumably realizing the same kind. For each problem, I will show that the underlying brain state/process types can be *relevantly* different, as measured by ϵ-isometry between the corresponding vector sets.

Binary Classification Under the assumption that our input images contain either cats or dogs, and cats always exhibit property p_c, and dogs always exhibit property p_d,

[7] Li et al. (2023) flatten voxel representations and use Gaussian smoothing to extract word-level signals. In such a vector space, however limited and approximative, we can define a number of psychological kinds, e.g., analogical inference, inflection, and passive alteration.

here are two ways of making similar decisions: (a) Predict something is a cat unless p_d, and (b) predict something is a dog unless p_c. Both of these algorithms can be implemented in deep neural networks. Assume our images are two-pixel images, our cat images are of the form $\langle 1, 0\rangle$, and our dog images of the form $\langle 0, 1\rangle$. Our first input feature is thus p_c; our second is p_d. Construct two simple perceptrons with two features and a bias term. One has weights $\langle 1, 0\rangle$, the other $\langle 0, 1\rangle$. Both have bias terms $b = -.1$. The first network will predict dog unless p_c. Our second network will predict cat unless p_d. Both networks implement the same function (or equivalent functions), but using different algorithms. Is this an example of MR? The two implementations of cat-dog discrimination are not generally ϵ-isometric for reasonable values of ϵ. To see this, consider how the perceptrons encode cats and dogs. The first model encodes cats as $\langle 1, 0\rangle$ and dogs as $\langle 0, 0\rangle$. The second model encodes cats as $\langle 0, 0\rangle$ and dogs as $\langle 0, 1\rangle$. The cross-distance class is thus always 1. Unless ϵ is set to more than ϵ, which would render the metric ineffective, the two models are therefore not ϵ-isometric. If vector spaces are normalized to unit length, we can generally assume $\epsilon \in [0, 1[$.

Division Ethnological studies have explored the use of different division algorithms among cultures, or even among high school students in the same high school. These algorithms are interesting, because they are different, but map the same input to the same output. The functions are thus equivalent, but different: Two division algorithms have the same domain and the same range, and for each element of the domain, the two algorithms yield the same result. Why are these algorithms not relevantly similar brain state/process types under ϵ-isometry? Consider two algorithms for division: (a) The so-called scaffold algorithm successively subtracts multiples of the divisor from the dividend. Say the problem is 52/4. We first subtract 4×5 from the 52, for example. This leaves us with 32. Seeing we still have a large remainder, we subtract 4×5 again, leaving us with 12. We see $12 = 4 \times 3$. The resulting quotient is therefore $5+5+3 = 13$. The key to the scaffold algorithm is our ability to estimate good multiples. (b) The standard algorithm substitutes the random search over possible multiples with a pass—left-to-right—over the digits of the dividend. We initially subtract 4×10, because we can only subtract 4 from 5 once. Because 5 is the second digit of 52, 1 has the value of 10, which means we subtract 4×10. This leaves us with 12 and a quotient of $10 + 3 = 13$. We see the standard algorithm is more efficient than scaffold algorithm with random search. What is important for our purposes is to see how a random search over multiples can produce arbitrarily long derivations as dividend grow to infinitude. This, in turn, will make the internal representations of these derivations, say, in a neural network, arbitrarily different from the representations using the standard algorithm. Two networks relying on the two algorithms will therefore not be ϵ-isometric.

Sorting In computer science, there are many so-called sorting algorithms, i.e., algorithms that put elements of a list into an order. Examples include insertion sort, merge sort, bubble sort, quick sort, and bucket sort, but there are more. Now is sorting a psychological kind, i.e., a mind state type or a mind process type? Why are two sorting algorithms not ϵ-isometric? Consider merge sort and quick sort, for example. Quick sort is faster than merge sort on small arrays, but the speed of merge sort depends less heavily on list length. Specifically, merge sort operates in $\mathcal{O}(n \log n)$, i.e., the time it takes to sort a list of n elements is less than quadratic in n. Quick sort, in contrast, is quadratic, i.e., in $\mathcal{O}(n^2)$. This means again that as n grows toward infini-

tude, the derivations from merge sort and quick sort will become arbitrarily different. Two networks implementing the two algorithms will thus exhibit arbitrarily different derivation vectors; therefore, the vector spaces are not ϵ-isometric.

For cat-dog discrimination, division, and sorting, different algorithms lead to relevantly different brain state/process types through the lens of ϵ-isometry. We still need to decide whether the different algorithms for cat-dog discrimination, division, and sorting are relevantly similar and instantiate the same kinds. I shall have little to say about whether cat-dog discrimination, division, and sorting are in fact *psychological* kinds—except that I see no good reason why they should not belong to the same class of kinds that emotions and color belong to—but focus on whether the different brain state/process types realize the same kinds.

We commonly rely on linguistic, behavioral, or experimental evidence when we establish psychological kind differences. Linguistic evidence is perhaps only relevant for studying folk taxonomies, but etymology has been evoked by notable philosophers in the past. Neither cat-dog discrimination, division, or sorting seems a good candidate for kind-splitting on linguistic grounds, though, and I will only discuss behavioral and experimental demarcation criteria.

Behavioral and experimental demarcation criteria *could*, in theory, establish kind differences. Now what would such criteria look like? What are the behavioral demarcation criteria for cat-dog discrimination, division, and sorting? For division and sorting, it is easy to see that implementations of different algorithms will not be guaranteed to lead to different answers. If someone relies on quick sort, they will return the correct answer, even for long lists, if they follow the algorithm slavishly. Humans are error-prone and may suffer from fatigue effects, rendering results more uncertain for longer lists, but this is no solid demarcation criteria for psychological kinds. Computational efficiency may be, however. Invariably, the time it takes implementations of quick sort to sort longer lists will grow significantly faster than the time it takes implementations of quick sort to do the same. This is, potentially, a demarcation criterion for splitting the psychological kind of sorting into two distinct psychological kinds.

Computational speed will not be a reliable way of distinguishing between our two algorithms for cat-dog discrimination. Simply presenting our implementations with hard-to-classify examples may help. If the two implementations are presented with an average image of $\langle 0.5, 0.5 \rangle$ (half-cat-half-dog), the two implementations will return different predictions. Our first perceptron will predict cat: the second dog. We can use such adversarial examples to split cat-dog discrimination into distinct kinds.

Consider next how experimental demarcation principles are also, in theory, possible. People who rely on different division strategies may, for example, exhibit significantly different skin conductance responses, as dividends grow toward infinitude. Or they may be more or less sensitive to intoxication. Or their neural activation may look different through the lens of modern brain imaging techniques. All of the above could, in theory, lead to operational demarcation principles that would facilitate kind splitting. If brain imaging was the preferred strategy among psychologists, we could again use ϵ-isometry (or representational similarity analysis) to establish our demarcation criteria. On a sample of division problems, for example, subjects will produce a set of (sets of) brain imaging vectors, and we could ask if two sets of subjects relying on different division algorithms exhibit vector spaces that are ϵ-isometric or not.

5 Concluding Remarks

The literature on MR is considerable. This short article does not reflect the full complexity of the debate, but suggests that much of it relies on *Bookof Sand* arguments. In the absence of demarcation criteria, such arguments successfully refute MR, but this is as much a problem for identity theory as it is for MR. For in the absence of demarcation criteria, identity theory is not falsifiable. I have discussed possible strategies for making identity theory (and MR) falsifiable. Brain state/process types can be differentiated by properly relaxed isometry measures. I leave open what is the best demarcation principle for psychological kinds, but suggest that behavioral and experimental criteria, e.g., statistical tests over variables such as response time, skin conductance, or brain imaging, are *possibly possible*, at least for some kinds.

Acknowledgements Thanks to the anonymous reviewers for helpful advice and detailed feedback.

Funding Open access funding provided by Copenhagen University.

Declarations

Conflict of Interest The author declares no competing interests.

References

Bechtel, W. (1999). Multiple realizability revisited: Linking cognitive and neural states. *Philosophy of Science, 66*(2), 175–207.

Boyd, R. (1989). What realism implies and what it does not. *Dialectica, 43*(1–2), 5–29.

Ekman, P., Levenson, R. W., & Friesen, W. V. (1983). Autonomic nervous system activity distinguishes among emotions. *Science, 221*(4616), 1208–10.

Figdor, C. (2010). Neuroscience and the multiple realization of cognitive functions. *Philosophy of Science, 77*(3), 419–456.

Kim, J. (1992). Multiple realization and the metaphysics of reduction. *Philosophy and Phenomenological Research, 52*(1), 1–26.

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis - Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience, 2*, 4.

Li, J., Karamolegkou, A., Kementchedjhieva, Y., Abdou, M., Lehmann, S., & Søgaard, A. (2023). Structural similarities between language models and neural response measurements. In *NeurIPS 2023 workshop on symmetry and geometry in neural representations*.

Michel, M. (2019). Fish and microchips: On fish pain and multiple realization. *Philosophical Studies, 176*(9), 2411–2428.

Polger, T. W. (2009). Evaluating the evidence for multiple realization. *Synthese, 167*(3), 457–472.

Polger, T. W. (2009). Evaluating the evidence for multiple realization. *Synthese, 167*(3), 457–472.

Premack, D. (2007). Human and animal cognition: Continuity and discontinuity. *Proceedings of the national academy of sciences, 104*(35), 13861–13867.
Putnam, H. (1960). Minds and machines. In S. Hook (Ed.), *Dimensions of Minds* (pp. 138–164). New York, USA: New York University Press.
Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience, 27*(169–92), 02.
Shapiro, L. A (2004). *The mind incarnate*. A Bradford book: MIT Press.
Sharma, J., Angelucci, A., & Sur, M. (2000). Induction of visual orientation modules in auditory cortex. *Nature, 404*, 841–847.
Shea, N. (2018). *Representation in cognitive science*. Oxford University Press.